ITEM POOL AND ITEM DEVELOPMENT TRAINING PACKAGE

CCSSO/SCASS ARTS EDUCATION ASSESSMENT CONSORTIUM

Updated by Vicki Fredrick in May, 2007 for the
CCSSO/SCASS Arts Education Assessment Consortium

Acknowledgments

In addition to the contributions of the SCASS Arts Education Assessment Consortium members, we would also like to recognize the members of the Comprehensive Assessment Systems (CAS) SCASS group and the SCASS Comprehensive Social Studies Assessment Project (CSSAP) for developing some of the materials used in this training package.

The photograph on page 34 is used by permission of Phil Graneto, Set & Costume Design for *Servant of Two Masters*; Rowan University, Glassboro, NJ

The scene design sketch on page 37 is used by permission of Bart Healy, Design Rendering; Rowan University, Glassboro, NJ

Table of Contents

This package contains a manual and PowerPoint presentation describing the process for developing selected-response multiple-choice items and constructed-response items, with and without stimulus materials, for the SCASS Arts Item Pool and handouts for use in the item development process.

Introduction

This manual is intended for use by state SCASS member states in training the states' SCASS Arts Item Pool item developers. The training manual describes a step-by-step process for drafting items and for critiquing items during and after the development process. The manual also includes handouts that can be reproduced for item development training and item writing, and it tells how/when to use each handout. This document is not intended to cover every possible detail of item development, but it does attempt to cover most of the major issues to be aware of in developing high quality multiple-choice selected-response items and constructed-response items with preliminary scoring criteria.

DESIGN

This Training Manual contains the following Five Sections.

SECTION I: ITEM DEVELOPMENT FOR THE ITEM POOL—TYPES OF ITEMS & COPYRIGHT OWNERSHIP

This one-page section provides a brief description of selected-response and constructed-response items and the structure of the Item Pool, and addresses release of copyright by item developers. Page 2.

SECTION II: DEVELOPING MULTIPLE-CHOICE SELECTED-RESPONSE ITEMS

This section of the document describes what a selected-response item is, lists various types of selected-response items, gives a brief explanation of why multiple-choice selected-response items are being developed for the Item Pool, defines the parts of a multiple-choice item, tells how/when to use each handout, describes a step-by-step process for drafting multiple-choice items and for critiquing items during and after the development process. The example items included are not arts items; they are drawn from areas of common knowledge so that anyone from any discipline can understand some of the common problems in these types of items and some suggested remedies for those problems, without having to have expertise in a particular discipline. Page 3.

SECTION III: DEVELOPING MULTIPLE-CHOICE SELECTED-RESPONSE ITEMS WITH STIMULUS MATERIALS

This section of the document describes what types of stimulus materials are compatible with the SCASS Arts Item Pool, important considerations for selecting stimulus materials for inclusion with items submitted to the Pool, and additional concerns related to writing multiple-choice selected-response items with stimulus materials. Example items in this section are arts items. Page 18.

SECTION IV: DEVELOPING CONSTRUCTED-RESPONSE ITEMS, WITH OR WITHOUT STIMULUS MATERIALS

This section of the document describes a process for developing constructed-response items, with or without stimulus materials, for the Item Pool. Development of preliminary scoring criteria is also addressed. Example items and scoring criteria in this section are arts items. Page 24.

SECTION V: HANDOUTS AND WORKSHEETS/CHECKLISTS (SEE PAGE 41 FOR COMPLETE LIST)

Universal Design and Sensitivity Considerations

Item Needs Summary for Each Discipline—(**Dance, Music, Theatre, & Visual Arts**)

National Standards for Each Discipline—(**Dance, Music, Theatre, & Visual Arts**)

Style Guides for Item Formats (multiple-choice and constructed-response items)

Worksheets for Creating Items (multiple-choice and constructed-response items)

Checklists for Creating Items/Stimuli (multiple-choice & constructed-response items)

Assignment of Copyright

Release Form

Section I

ITEM DEVELOPMENT FOR THE ITEM POOL—TYPES OF ITEMS AND COPYRIGHT OWNERSHIP

TYPES OF ITEMS—SELECTED-RESPONSE AND CONSTRUCTED-RESPONSE

In the broadest definition of item types, there are two general categories: (1) **selected-response** items in which students select the correct or best answer from the given options, and (2) **constructed-response** items in which students provide their own answers or responses. Selected-response items include multiple-choice, true/false, and matching items. Constructed-response items include fill-in-the-blank, short-answer, on-demand performance, extended-response, portfolio, etc.

Initially, item development for the SCASS Arts Item Pool focused on selected-response items. However, the long-range goal of Item Pool development was, and continues to be, to provide member states with a collection of selected-response and constructed-response raw items from which to choose. These untried items could then be refined and field tested by member states to meet or supplement their professional development and/or assessment needs. The purpose of an assessment should determine which types of items to include in an assessment instrument. It will be the responsibility of the user-states to make those decisions and to conduct appropriate trials of items that they choose to use from the Item Pool.

The major advantage to using multiple-choice and other varieties of selected-response items is that such items can be used to assess a broad range of student knowledge and understanding of many things in a short time; the major disadvantage of selected-response items is that they do not allow students to go into depth or to create their own responses as constructed-response items can do. Within the practical limitations of time and cost, if possible, it is usually best to include a balance of selected-response and constructed-response items in an assessment.

COPYRIGHT OWNERSHIP OF ITEMS IN THE ITEM POOL

All item developers are required to sign a **Release Form** and an **Assignment of Copyright** form before developing items to be entered into the Item Pool Review Process. Those forms are included at the end of this training manual. Item developers can choose to have their names listed as contributors, or not, by checking one of the boxes on the Release Form. It should be noted that almost all items undergo multiple revisions and changes during the development and review process, and several people usually contribute to the final version of most items.

Section II

DEVELOPING MULTIPLE-CHOICE SELECTED-RESPONSE ITEMS

SELECTED-RESPONSE ITEMS FOR THE SCASS ARTS ITEM POOL

As noted in Section I on the previous page, selected-response items include a variety of formats in which students are directed to select the best or correct response from among given options. For the SCASS Arts Item Pool, the only types of selected-response items that are included at this time are multiple-choice items.

The multiple-choice item format consists of a stem, which is either in the form of a question to be answered or a statement to be completed, and possible options from which the student can select the correct response. The SCASS Arts Item Pool format requires that only one option can be selected by the student. There is only one correct answer per item. The correct answer is called the key. The incorrect options are called distractors (or foils). For the SCASS Arts Item Pool, each item will have a stem and four options (the key and three distractors). Following is an example that shows the parts of a multiple-choice item:

	Stem	Which of these rivers forms part of the boundary between the United States and Canada?
Options	**Distractor**	A. the Rio Grande River
	Key	B. the Saint Lawrence River
	Distractor	C. the Wisconsin River
	Distractor	D. the Mississippi River

In creating multiple-choice and other types of selected-response items, the tendency is to create items that tend to focus on lower-level factual information unless care is taken to elevate the thinking required to answer the items. For the SCASS Arts Item Pool, item developers are strongly encouraged to focus on developing items that go beyond recall and basic knowledge items. Although some items may focus on recall or knowledge of important facts, major emphasis should be placed on also developing items that measure as many higher levels of understanding of important arts concepts as possible, such as:

- Comprehension
- Application
- Interpretation
- Analysis
- Evaluation

Items should also be developed to be appropriate for the greatest number and broadest audience of students. Throughout the item development and critiquing process, item writers must, of course, be aware of issues of sensitivity when developing items. See the **Universal Design and Sensitivity Considerations** handout included with this manual for a list of some issues to keep in mind. States that use items from the Item Pool will have various procedures for reviewing items for these issues and others that may be unique to the user state.

CREATING A MULTIPLE-CHOICE ITEM: A STEP-BY-STEP APPROACH

Getting Started

The following approach is provided for the purpose of developing music, theatre, dance, and visual arts multiple-choice items to add to and enhance, in a systematic way, the breadth and depth of the items in the SCASS Arts Item Pool. To help direct the item development process, a current summary of items in the Item Pool for each discipline can be obtained directly from the Item Pool's reports.

Included with this Training Packet, is an example of a SCASS Arts **Item Needs Summary** handout for each of the four arts disciplines:

- music,
- theatre,
- dance, and
- visual arts.

The example **Item Needs Summary** for each discipline included in this manual is based on information available in May 2003, and the number of items needed for each discipline and standard changes as items are added to the pool, so, when item developers begin to write items, they should have a copy of the current information from the Item Pool reports. Item developers should begin by reviewing the current Item Needs Summary for the discipline for which they will be developing items. The Item Needs Summary lists the **National Standards**, grade levels, key descriptors, and item levels (knowledge/recall, comprehension, interpretation, application, analysis, synthesis, and/or evaluation) for which the greatest needs currently exist.

After noting where the greatest needs are for the discipline for which they will be developing items, item developers should then choose a National Standard for which there are few, if any, items. All items for the Item Pool must be written to measure an important aspect of a National Standard. The National Standards handout included in this training manual lists a one-line descriptive statement for each standard in each discipline. For the SCASS Arts Item Pool, these one-line statements should be used to indicate which standard an item is designed to measure. To make the Item Pool useful for a variety of member states, there must be a common reference point. Therefore, the National Standards statements listed will be used for that purpose. Ultimately, if a member state wishes to use items from the Item Pool, that state can match items to state standards or use whatever categorization system that is appropriate for the state's use of the items. Having the National Standards as a reference point will assist in that task.

Regarding grade level, item writers should concentrate on developing items for the grades with which they are most familiar, as items at all grade levels are needed. If an item developer is familiar with multiple grade levels, then he/she should concentrate first on developing items at grade levels where the greatest needs exist. However, in general, multiple-choice items usually are not recommended or often used for kindergarten through grade 2 children, and use of multiple-choice items for grade 3 children is fairly uncommon. Obviously, great care and expertise regarding the language development levels of children at these grade levels is required when creating items because the life experiences, vocabulary, and understanding of written words are still quite limited in very young children. Items in the Pool that are designated as appropriate for kindergarten-grade 2 were developed with the assumption that the test administrator would read these items aloud to students. Items at other grade levels may also be read aloud, depending on circumstances. Such decisions are to be made by the user states.

If an item is suitable for more than one grade (e.g., grades 6-8), this grade range may be indicated. Users of the Pool will make the ultimate decision regarding the grade(s) for which an item is/are most appropriate, as curriculum varies from state to state. Grade levels in the Item Pool are meant as guides to the user, not absolutes.

DEVELOPING MULTIPLE-CHOICE ITEMS:

It is a challenging task to craft an item, consisting of about five short lines of words, so that all facets of the item are working well together, and the item is measuring something that is important to know—all at the same time. To meet the challenge of crafting high quality items for the SCASS Arts Item Pool, a recommended procedure is presented in this training manual.

The step-by-step item development procedure described in this document is based on procedures that have been followed for many years by many states and test development companies. It has been proven to work well when developing multiple-choice items no matter what the subject area or discipline. States participating in the SCASS Arts Item Pool development process may choose to make some adaptations in the process to fit within the procedures of the member state; however, any adaptations should still result in the development of items that meet the needs and criteria of the SCASS Arts Item Pool.

It is recommended that item developers work in small groups or with a partner, rather than in isolation, even if they already have had extensive item development experience. Sharing of ideas and bouncing suggestions back and forth among colleagues, at the item creation level and when revising items, usually results in higher quality items with clearer focus.

Typically, a high quality item doesn't happen in the first draft or, often, not even in the second or third draft. Numerous reviews and revisions are almost always necessary to bring an item to a level of quality that will make it a good, useful item. Item developers come to understand that revision, revision, and more revision is a large part of the normal process of item development. It is important for new item developers to be aware that self-critiquing and peer-critiquing are normal and expected parts of developing high quality items.

Item developers should each have a copy of the SCASS Arts Item Pool **Style Guide for Multiple-Choice Selected-Response Items** included in this training manual to follow during item development. It is very important that all item developers follow this Style Guide so that all items will be in the same format, no matter which state develops the items. If the enclosed Style Guide varies from what the item writer may be accustomed to from other item development projects, the item writer must follow the SCASS Arts Item Pool Style Guide.

Keeping all of the previously stated information in mind, item developers should use the **Worksheet for Creating a Multiple-Choice Selected-Response Item** included in this training manual for rough-drafting items (or create their own form containing the same information) and follow the step-by-step procedure below for constructing each item:

Note: For entry into the on-line Item Pool review process, each item must be given a short title to identify it. It may be easiest to name the item after it has been completed. The Item Title should be written on the Worksheet.

STEP 1: National Standard & Grade Level

Choose a National Standard and grade level upon which to focus, as described on the previous page. (See the "Getting Started" section.) Enter the National Standard statement and grade level(s) that the item will address on the "Worksheet for Creating a Multiple-Choice Item."

STEP 2: Information/Concept Focus of Item

Decide what important information or concept the item will address. Enter this information on the "Creating a Multiple-Choice Item" worksheet. What will this item attempt to find out from a student about this important information or concept? Be as clear and concise about what the item will measure as is possible at this point; it probably will be necessary, as the item progresses, to keep refining and honing its focus.

STEP 3: The Stem

Draft an item stem that addresses the concept or information identified in step two, keeping in mind that items measuring levels of thinking higher than recall/knowledge are needed most. Enter the draft item

stem on the worksheet, directing the item toward an intended level of thinking (comprehension, application, interpretation, analysis, or evaluation).

The stem is usually a question, but the stem can be in the form of a statement to be completed (see "Style Guide for Multiple-Choice Selected-Response Items"). No matter which format the stem uses, it must be clear to students, from the stem alone (i.e., without the options), what is being asked. If it won't become clear what the question is asking until one sees the options, then it is not an acceptable stem.

Refine the wording of the stem to be as clear and concise as possible, paying attention to appropriate wording for the grade level. The natural tendency for adult item developers is to write at an adult level, so careful attention must be paid to appropriateness for students, especially at the elementary level. Typically, the stem is no longer than a line or two.

Tip: If it is very difficult to get the wording of the stem just right at this point in the drafting stage, it might help to write the correct answer (key), then return to the stem for further refinement.

STEP 4: The Key

Draft the key (the correct answer) and enter it on the worksheet. The key must correctly answer the question asked in the stem or correctly complete the statement in the stem. The answer must be a complete and accurate answer to the question asked in the stem; partial answers make an item unnecessarily confusing to students and, therefore, are not acceptable.

Refine the wording of the key to be as clear and concise as possible, using language appropriate to the grade level. Usually, the key is no more than a line or two long.

Carefully consider the stem and key together. Taken together as a unit, the stem and the key must address the intended information or concept at the item level intended. At this point, some refinement or reworking of the stem and key may be necessary, taking care not to lose sight of the intended purpose of the item. It may even be necessary to try a completely different approach to measuring the desired concept. Adjustments should be made in the stem and/or key until the stem/key match is satisfactory.

STEP 5: The Distractors

Draft the distractors (the three incorrect answers) and enter them on the worksheet. The distractors must be completely incorrect, while still being plausible responses to the stem. A distractor that is partly correct and partly incorrect is confusing and, therefore, not acceptable.

Each distractor should be distinct from the other options; overlapping information in distractors is confusing and, therefore, not acceptable.

Distractors should attempt to reflect actual errors in understanding that students might have about the concept, but they should not be tricky. There should be some reasonable likelihood that a student who does not know the correct answer might logically pick the distractor (hence the term, distractor). If a distractor is so implausible that no one will choose it, then it is not useful in gathering any information about students' misconceptions. Also, the set of distractors should not incorporate multiple types of errors in misunderstanding because this would be confusing and make it difficult for teachers to determine what led to the student's choosing a particular distractor.

Refine the wording of the distractors, using grade-level appropriate language. The distractors should be parallel to the key: similar in length, similar in grammatical structure, similar in appearance (e.g., graphic cues such as capital letters, hyphens, apostrophes, etc.), and similar in concept.

Carefully consider the key and distractors together, and consider all the options as they function with the stem. They should function well as a whole item—addressing the intended National Standard, the intended information or concept, the intended item level, etc. Some additional refinement may be necessary in the stem, the key, and/or distractors.

STEP 6: Level

Confirm the level of the item—knowledge, comprehension, or higher (application, interpretation, analysis, or evaluation) and indicate the level on the worksheet. Sometimes, during the process of refinement, the item level is inadvertently or intentionally transformed from the level that was originally

intended. A change in item level during the drafting process is okay when the change results in a better item. However, it is important to re-check the level and indicate the correct level on the "Creating a Multiple-Choice Item" worksheet. Make the most reasonable judgment about its level that is possible, keeping in mind that the main distinctions we are concerned about are knowledge vs. higher level thinking. If it is clear that the item measures thinking at a level higher than the knowledge level, but which of the higher levels is being measured is difficult to distinguish, then item developers should use their best professional judgment as to the appropriate level.

STEP 7: Evaluate, Revise, and Finalize

Use the Checklist for Creating and Evaluating Multiple-Choice Items included in this training manual to evaluate the rough-draft items regarding the significance of the item and the technical qualities of the item. The checklist is intended to assist in evaluating, revising, and finalizing items constructed for the Item Pool. The goal is to have items as refined as possible before they are entered into the on-line review process. Once all the boxes on the form can be checked (indicating "okay") then the item is ready to be presented to the item developer's State Coach or the person designated to enter items into the Item Pool's on-line review process.

The section, "Explanation of the Checklist for Creating and Evaluating Multiple-choice Items," which begins on the following page is intended to supplement and explain the criteria listed on the "Checklist for Creating and Evaluating Multiple-Choice Items". In some instances, sample problem items and suggested improvements are shown.

In addition to using the checklist and the explanation of the checklist during the drafting process, they should also be used to make sure an item is ready to present to the State Coach for possible entry into the Item Pool review process. (The Item Pool process is described in separate documentation.)

EXPLANATION OF THE CHECKLIST FOR CREATING AND EVALUATING MULTIPLE-CHOICE ITEMS:

The ***Checklist for Creating and Evaluating Multiple-Choice Items*** included in this training manual should be used during the process of developing an item to guide its construction and after an item is developed to ensure that the item meets the criteria for consideration for the SCASS Arts Item Pool. Before submitting the item to the State Coach for review and possible entry into the Item Pool Holding Area, the item developer(s) should check the item for significance and technical qualities using the checklist. For each criterion on the checklist, the item developer should check the box (indicating "yes") if the item meets that criterion. If the item does not meet one or more of the criteria on the checklist, the item should be revised until it meets all criteria. If it cannot be revised to meet all criteria, the item should not be submitted for review. If the checklist is used throughout the item development process, there should be very few items that will not meet the criteria at the final check stage.

The following information describes in more detail the criteria on the checklist. In some cases, an example of an item that is flawed and suggestions regarding how to revise the item are given.

SIGNIFICANCE OF THE ITEM

1. The item relates directly to a National Standard.

Since the SCASS Arts Item Pool uses the National Standards as a common reference point for all member states, it is critical that each and every item developed/considered for the Item Pool relates clearly and directly to a National Standard. Item developers should use the one-line, general statements listed on the "National Standards for Item Entry Form" that indicate each of the National Standards. Item developers should not be concerned about pin-pointing exact subpoints given in the complete version of the National Standards, but rather they should focus on these general descriptive statements for item categorization. The general statement of the standard addressed will be indicated on the Item Pool entry form when the item is considered ready for entry. These statements will serve as a search guide to users of the Item Pool and will enable states to match items from the Item Pool to their own state standards, if they choose to do so. Item developers should be aware that most, but not all, of the standards are suited to being measured by selected-response items, at least to some degree.

2. The item deals with information or a concept of importance, not with trivia.

Using the selected National Standard as a guide should help item developers know what SCASS Arts considers important areas in each arts discipline. However, the item content must focus on important information/concepts within the National Standard. If one can answer the following question affirmatively, then the item is acceptable: If a student knows the answer to this question will that tell us something important about what the student knows, understands, etc.?

Problem item:

How many lakes make up the Great Lakes?

A. two
B. three
C. four
D. five (Key)

This item focuses on very low level recall, and, as such, reveals little about what the test-taker knows about the Great Lakes or their significance. Sometimes, although not always, low level recall items can be reworked to require higher level thinking.

Suggested improvement:

Why are the Great Lakes most likely of increasing interest to the people of southern California?

A. because they want to sail their boats on the Great Lakes in the summer
B. because they view the Great Lakes as a potential source for fresh water (Key)
C. because they need fish from the Great Lakes in order to have enough food
D. because they would like to harvest ice from the Great Lakes in the winter

The revised item requires deeper levels of understanding, analysis, and comparing/contrasting. If a student responds correctly, it will indicate that he/she understands the connection between the fact that the Great Lakes are a substantial source of fresh water and that fresh water is in increasing demand. This is an important concept and reveals much more about a student's understanding of natural resources or physical features of geography than simply being able to recall how many lakes make up the Great Lakes.

To answer the question, it would be helpful to understand that the Great Lakes are large bodies of fresh water, rather then salt water, which is abundant along the coast of California. Further, the student would benefit from knowing that southern California has such rapidly increasing needs for water that even a faraway source such as the Great Lakes might be attractive. Comparing and contrasting climate and population growth of California and the Midwest would be helpful in answering this question.

The three distractors are plausible because recreation, fishing, and ice harvesting are all possible uses of the Great Lakes, but the need for places to sail boats, obtain fish to eat, or get ice can be satisfied locally in California, whereas, the need for more fresh water cannot. The repetition of the words "because they" or even the use of the pronoun "they" in each option can be debated. Generally, repetition should be avoided unless the repetition makes the options clearer, and any pronouns used must have clear referents. These are judgments to be made be the item developers and reviewers.

3. The item is tightly focused so that it will be clear why a student gets it right or wrong.

A multiple-choice item should try to measure a student's understanding of only one concept or important piece of knowledge. It is easy to inadvertently introduce additional concepts through the distractors, so a careful critique of what decisions a student will have to make when sorting through all four options is necessary. If an item is too ambitious or requires a student to make choices between different concepts presented in a single item, then, it quite likely will not be clear why a student chose the answer that he/she did. If the reason for selection is not clear, it will not be clear whether or not the student understands the intended concept or knowledge measured by the item.

Problem item:

Which of these cities is the largest city located on Lake Michigan?

A. Chicago (Key)
B. Duluth
C. Milwaukee
D. Detroit

In addition to focusing on low level recall, this item is mixing two pieces of information: (1) which of the cities listed is the largest, and (2) which is located on Lake Michigan. If a student answers this item incorrectly, will it be because he/she didn't know which of the cities was the largest or which is located on Lake Michigan? If it is assumed that this information is important enough to know, the item should be reworked to focus on one piece of information.

Suggested improvement:

Which of these cities is the largest city located on Lake Michigan?

A. Chicago (Key)
B. Manitowoc
C. Milwaukee
D. Racine

The options in this case are all cities located on Lake Michigan, so the task in this item is to identify which of these is the largest city.

or:

Which of these large cities is located on Lake Michigan?

A. Chicago (Key)
B. Detroit
C. Minneapolis
D. Montreal

The options in this case are all large mid-continent cities, but only one is located on Lake Michigan, so the task is to identify which of the given cities is located on Lake Michigan.

Because this item is a simple recall level item, it was fairly easy to see that in its original form, it was measuring more than one piece of knowledge. In higher level items, often it is not as easy to see, even though it can occur inadvertently. Diligence must be exercised in constructing items at any level.

4. If beyond recall or knowledge level, the item asks students to use higher-order thinking skills such as:

- Comprehending an important concept or information
- Applying a concept/skill to a different situation
- Making an interpretation/drawing a conclusion
- Comparing/contrasting concepts or information
- Analyzing and/or evaluating a claim or generalization

The emphasis for the SCASS Arts Item Pool is on developing items at levels higher than recall or knowledge. For purposes of this item pool, it is less important to tightly categorize items as to which of the higher levels an item is incorporating, than it is to distinguish between items at the recall/knowledge level vs. those at some higher level. Therefore, if it is difficult to draw fine lines between the higher levels, it is sufficient to use one's best professional judgment when classifying an item. For example, there might be analysis involved when comparing two concepts, or there might be application involved when evaluating a claim, etc. In these cases, the item developer should identify what he/she believes is the most appropriate classification level.

Problem item:

Which of these large cities is located on Lake Michigan?

A. Chicago (Key)
B. Detroit
C. Minneapolis
D. Montreal

Although this item is technically adequate, it is focusing on low level recall. From a student's answer, all that will be evident is that the student knows or doesn't know that, of these cities, only Chicago is located on Lake Michigan. Perhaps this item could be reworked to ask about a more significant concept related to the location of Chicago.

Suggested improvement:

Why did Chicago's location on Lake Michigan at the mouth of the Chicago River contribute to its rapid growth in the early 1800's?

A. Year-round mild climate conditions near the water made people want to live there.
B. The area was protected by the water from nearly all types of natural disasters.
C. Most long-distance transportation of goods and people was by water. (Key)
D. Land was less expensive near the water than it was farther away from the lake.

The revised item focuses on a more important concept—the concept of transportation's effects on settlement and growth. In pre-railroad days, the predominant mode of transportation in North America (and elsewhere around the world) was via a network of waterways. Cities like Chicago were located at points along these waterways where goods would be traded and/or transferred from canoes or small boats to larger vessels that could cross the larger bodies of water and transport goods to the eastern population centers.

5. The item addresses sensitivity issues and the concept of "universal design" for the broadest audience.

A final check of the item for problems related to sensitivity issues that could produce a biased item is important. This would include a check to be sure the item has been written with the broadest possible audience of students in mind. The item should not cause unnecessary challenges to students with disabilities, English language learners, or others. The item should not present negative stereotypes toward any group, such as those groups listed on the ***Universal Design & Sensitivity Considerations*** handout. If there is any doubt that the item might be offensive to a particular group, favor one group over another, or exclude some students unnecessarily, the item should be reworked.

TECHNICAL QUALITIES OF THE ITEM:

1. The stem is a complete question or statement; the intent of the stem is clear without reading the options.

Make sure that each stem asks a clear question or requires students to complete a coherent statement. The stem should inform students exactly about what they are expected to know about the content. A simple method of testing for clear intent in a stem is to cover all four options and see if it is possible to give a correct response when looking only at the stem. With the options covered, is it clear what is being asked of the student? If not, the stem should be reworked to make it clear without depending on the options for clarification of intent. Sometimes, this involves reworking a distractor(s), too.

Problem item:

George Washington was the American who

A. led the colonial army against the British. (Key)
B. wrote the Declaration of Independence.
C. fired the first shot in the Revolutionary War.
D. organized the Boston Tea Party

Notice that when the four options are covered from view, it is impossible to tell what the question is asking about George Washington.

Suggested improvement:

What was George Washington's most important role during the American Revolution?

A. He led the colonial army against the British. (Key)
B. He wrote the Declaration of Independence.
C. He fired the first shot at Concord.
D. He organized the Boston Tea Party.

Note: The SCASS Arts "Style Guide for Multiple-Choice Selected-Response Items" states in point #1—"State the stem in question form if possible." Therefore, the stem has been re-written so that it is a question rather than a statement to be completed. The stem now makes it clear what the item wants to know about George Washington—his most important role during the American Revolution. Notice that inserting "the American Revolution" in the stem necessitated editing out "the Revolutionary War" from distractor "C" to lessen the possibility of confusion.

2. Information in the stem does not cue the key.

Information presented in the stimulus or stem should not provide a cue to the key. In the simplest terms, well-written items should require students to use relevant knowledge and skills, rather than test-wiseness skills, to determine the correct answers. While it is true that sometimes students can use some fairly sophisticated analytical skills to find cues to an answer, the focus of the assessment should be on the students' comprehension of specific content and conceptual knowledge, and items that provide unintended cues for test-wise students should be avoided.

Problem item:

Which of the following documents was written to guarantee certain rights to American citizens?

A. Emancipation Proclamation
B. Articles of Confederation
C. Bill of Rights (Key)
D. Declaration of Independence

The use of the word "rights" in the stem cues option C—Bill of Rights. Therefore, a student might pick the right answer without knowing anything about the options given. See the suggested improvement below.

Suggested Improvement:

Which of the following documents was intended to increase the likelihood that states would approve adoption of the United States Constitution?

A. Emancipation Proclamation
B. Articles of Confederation
C. Bill of Rights (Key)
D. Declaration of Independence

The revised item requires an understanding of why the Bill of Rights was incorporated into the Constitution, not what the Bill of Rights does for American citizens. In the process of removing the cue in the stem, the focus of the item was changed and the level of thinking was raised.

3. Negative stems are avoided.

Stems using negative words or prefixes, such as not, un, non, least, which of these doesn't, or all of the following except, etc., generally serve to complicate an item. Difficult-to-read stems are more likely to measure students' ability to decipher the question than their knowledge or skills. Usually, there are clearer ways to phrase the question. If the stem contains a negative, it should be reworked, if at all possible. In the process of reworking a negative stem, a distractor(s) might have to be reworked, also. In rare instances, it may not be possible to avoid the use of a negative in the stem. In these cases, extreme care is necessary in making the item clear to students.

4. There is one and only one clearly correct answer and the key is indicated.

There should be one correct answer indicated. The correct answer should be complete; that is, it should contain the whole answer, not just part of it. The distractors should be plausible responses, but they should be completely incorrect. Distractors that are partially correct and partially incorrect are not acceptable because they are confusing. If there is any doubt as to whether or not there is one and only one completely correct answer, the item should be reworked.

5. The position of the key is varied, while the order of the options is logical and appropriate.

Whenever possible, the correct responses should be randomly assigned as A, B, C, and D. Even though these items will be in an item pool, rather than on a test form, items with a variety of correct answers (A, B, C, and D positions) will aid those who might select items to include in an assessment instrument. On a test instrument, patterns of correct answers can provide shrewd students with an advantage unrelated to the focus of the test. If students get items correct for the wrong reason, then the test's validity comes into question.

Occasionally, the logical presentation of the options will demand a certain position for the correct answer, such as when the options are numbers, seasons, or some other type of response for which randomizing their order would only serve to confuse. For example, if the options are: two hours, three hours, five hours, and nine hours, it would be most logical to list them in either increasing or decreasing lengths of time, regardless of which is the correct answer. To randomize them would only confuse, not make a better item (e.g., five hours, nine hours, three hours, and two hours).

6. All distractors are plausible, yet incorrect.

Be sure that the distractors in multiple-choice items are plausible, yet incorrect. Writing good distractors is perhaps the most difficult part of developing multiple-choice items. Often an item writer will be able to write two distractors fairly easily, but will experience difficulty creating the last distractor. Because of this, sometimes an item writer will write a distractor that is a "give-away." Ideally, distractors should be common student misconceptions that are attractive to students who may only have partial understanding of the content. Although good distractors should seem plausible, care is needed to ensure that they are not so good that they are possible correct answers. Each multiple-choice item should have one and only one correct answer.

7. Overlapping options are avoided.

Make sure that information in the options does not overlap. Following this guideline also helps to avoid double-keyed items or to provide unintended cues to answering the item. Even if overlap affects only the distractors and the key is unaffected, test-wise students often rule out options without using the requisite knowledge of the content or skill. All overlap among options should be avoided.

Problem item:

When was slavery abolished in the United States?

A. after the American Revolution
B. after the War of 1812
C. after the Mexican War
D. after the Civil War (Key)

Because slavery was abolished after the Civil War, it is also true that slavery was abolished after the Mexican War, the War of 1812, and the American Revolution because the Civil War took place after all of those wars. This is an item with four possible correct answers.

Suggested improvement:

When was slavery abolished in the United States?

A. In 1787, by the U.S. Constitution
B. In 1814, by the War of 1812 Treaty
C. In 1848, by the Treaty of Guadalupe-Hidalgo
D. In 1865, by the 13th Amendment (Key)

There is now only one correct answer, and the options are distinct from one another—no overlapping time periods.

8. Options are parallel in concept, language structure, and appearance.

All options should be parallel. For example, if one option is a phrase, then the other options should also be phrases. Verb tense, use of adjectives and adverbs, use of plural/singular forms, and all other types of language usage should be parallel in the options and consistent with the stem. The concepts in all options should be parallel in type (e.g., all general concepts that are plausible, all specific details that are plausible, etc.). The appearance of the options should be parallel—use of proper nouns, use of possessives, use of commas, use of compound statements, etc. Here is an example and a suggested improvement.

Problem item:

Which is the best example of an actor improvising lines?

A. an actor memorizes the lines in the script and says them exactly as they are written
B. using movement and facial expression without speaking any lines
C. cue cards concealed from the audience
D. The actor created appropriate lines and delivered the lines as the scene developed. (Key)

This item has a number of problems with parallelism of various types.

Suggested improvement:

Which is the best example of an actor improvising lines?

A. The actor memorizes the lines in the script and says them exactly as they are written.
B. The actor uses movement and facial expression without speaking any lines.
C. The actor reads lines from cue cards that are concealed from audience members.
D. The actor creates appropriate lines and delivers the lines as the scene develops. (Key)

9. Options are of equal or nearly equal length.

There is a tendency for item writers to make the correct answer longer than the incorrect choices, often because qualifiers are used to make the key completely accurate. Adding qualifiers to the distractors or editing the correct option, while keeping all options parallel as described above, could help solve this problem. Remember that a distractor that is much longer or shorter than the others stands out and gives the “test wise” student a real or false clue to the correct answer.

10. Language usage and grammar in the stem and options are correct.

It would be helpful to have an expert in language usage and grammar review and edit items, especially if the item developers have questions about particular items they have written or if they feel less than confident about usage or grammar in particular items. Even if the content and concept of an item is excellent, a lapse in correct language usage can detract from all of the good qualities of the item.

11. The stem and options fit correctly/appropriately with each other, grammatically and conceptually.

After the item is considered "finished," it would be helpful to read the stem and all of the options one last time to make sure that not only all of the individual elements of the item are satisfactory, but also that all of the pieces hang together to make a whole item that is as coherent and smooth as possible. Sometimes, in fine-tuning the stem and/or the options, it is easy to lose sight of the whole item—the "can't see the forest for the trees" syndrome.

12. Complex formats in the stem, options, and overall item are avoided.

Avoid the use of complex formats or wording structures. There are rare occasions where complex formats can be useful, such as when testing for an understanding of multiple sources, causes, or influences. However, such formats add complexity to these items that is not related to the content being measured, and this will disadvantage some students. Generally, it is a good idea to present items as clearly and simply as possible so that there is little doubt that the item is measuring some relevant knowledge or skill. If an assessment uses a complex format, make every effort to keep the items as clear as possible and provide students with practice on such items prior to the test.

Problem Item:

Which of the following levels of government may levy taxes?

I. Local government
II. State government
III. Federal government

A. I and II only
B. I and III only
C. II and III only
D. I, II, and III (Key)

This item is very complicated. If it is important for students to know that all three levels of government in the U.S. can levy taxes, then perhaps the question can be asked more directly.

Suggested improvement:

In the United States, which of the following can local, state, and federal government do?

A. issue postage stamps
B. print money
C. declare war
D. levy taxes (Key)

The revised item requires students to distinguish a power that all three levels of government have from powers that only the federal government has. By using different distractors, the focus of the item could be changed.

13. Repetitive language in the stem and the options is avoided.

If certain words are repeated in each option, deleting those words or incorporating them in the stem can reduce the reading load of the item.

Problem item:

What did the Jamestown colonists eventually do to make their settlement financially successful?

A. The colonists began to build faster ships to cross the Atlantic Ocean.
B. The colonists began to grow a crop that they could export to England. (Key)
C. The colonists moved their settlement closer to Plymouth Colony.
D. The colonists began to sell manufactured goods to France.

The "colonists" are already mentioned in the stem, and they are the only people mentioned. Therefore, it is not necessary to repeat "The colonists" in each option. Also, in A, B, and D the phrase "began to" is used. To shorten the options and make them parallel, the same structure should be used in all four options—a simple past tense verb would suffice (see "Suggested improvement" below).

Suggested improvement:

What did the Jamestown colonists eventually do to make their settlement financially successful?

A. built faster ships to cross the Atlantic Ocean
B. grew a crop for export to England (Key)
C. sold manufactured goods to France
D. moved closer to Plymouth Colony

The wording in the revised item is much more streamlined and easier to read, as a result.

14. Superfluous wording in the stem and the options is avoided.

Keep the wording in the stem to the point. Overwriting or inserting irrelevant information into the stem adds unnecessary or unfair complexity to the item. Using an item stem as a "teachable moment" adds unnecessary complexity to the item. It is important to keep the grade level of the students in mind, also.

15. The item conforms to the SCASS Item Pool Style Guide for Multiple-Choice Selected-Response Items.

To standardize the item format, each item must follow the Item Pool Style Guide for Multiple-Choice Selected-Response Items.

16. The use of "all of the above" and "none of the above" as options are avoided.

Item writers sometimes use an "all of the above", "none of the above", or a similar statement as a distractor because of the difficulty of identifying a fourth option that is plausible, but incorrect. Sometimes, item writers use such statements as options because the particular test instrument designers allow the use of these statements as options. However, these types of options are not acceptable for the SCASS Arts Item Pool.

Offering "none of the above" as an option usually does not indicate what the student knows to be correct information. Rather, it shows only that the student knows that the other three options are not correct.

Also, if “all of the above” is offered as an option, and if a student can recognize that at least one of the other choices is incorrect, logically “all of the above” can be ruled out immediately as a correct answer.

17. Items are independent to the extent possible; distractors are not used in more than one item.

Items developed for the Item Pool may ultimately be used together on an assessment instrument, even though they will be entered as individual items into the Item Pool initially. Therefore, it is important for an item developer to avoid using the same distractors in different items. While it is impossible for item developers in one setting to know which distractors are being written by item developers in another setting, if each item developer or team of item developers avoids repeating distractors, the chance of different items using the same distractors will be minimized.

Note: If the item has stimulus materials, then the following two points will need to be addressed:

18. Any stimulus materials fit with the stem, are of high quality, and are adequately identified.

19. Any stimulus materials conform to the specifications of the Item Pool.

(For additional information regarding #18 and #19, refer to the section about selecting and using stimulus materials starting on page 18.)

Section III

DEVELOPING MULTIPLE-CHOICE SELECTED-RESPONSE ITEMS WITH STIMULUS MATERIALS

PURPOSE OF DEVELOPING MULTIPLE-CHOICE ITEMS WITH STIMULUS MATERIALS

As noted earlier in this training manual, multiple-choice items can be an efficient way to gather a significant amount of information regarding what students understand about a subject area. Therefore, the plan for the SCASS Arts Item Pool was and continues to be to build the Item Pool first with multiple-choice items, especially for grade levels and standards for which items were/are lacking, then to continue expanding and enhancing the Item Pool through including multiple-choice items with stimulus materials and constructed-response items, with and without stimulus materials. From the beginning, the emphasis for the SCASS Arts Item Pool has been on developing items measuring important concepts at levels higher than recall/ knowledge, such as items that measure students' ability to:

- Comprehend important concepts or information
- Apply a concept/skill to a different situation
- Make an interpretation/draw a conclusion
- Compare/contrast concepts or information
- Analyze a situation, object, or performance
- Evaluate a performance/object, claim or generalization
- And other higher level skills and concepts

In working toward the goals of the SCASS Item Pool, higher-level multiple-choice items without stimulus materials will continue to be developed, but the SCASS Arts Item Pool is now moving towards the next phase of development; that is, including multiple-choice items with stimulus materials and including constructed-response items. Development of multiple-choice items with stimulus materials is addressed in this section of this training manual, and constructed-response item development is addressed beginning on page 24.

Types of Stimulus Materials

For the SCASS Arts Item Pool at this time, still images, video images, text, music notation, and audio clips can be used with multiple-choice items and can be submitted through the on-line review system. Stimulus materials for SCASS Arts items should be selected from collections provided by SCASS Arts or from stimulus materials that have already been accepted into the Item Pool. Copyright clearance has already been negotiated and completed by SCASS Arts for stimulus materials in these collections.

Other copyright-free stimulus materials and collections may be considered on a case-by-case basis by the SCASS Arts Consortium. If an item incorporates stimulus material, the stimulus must be copyright-free and must be submitted along with the item into the on-line review process. The on-line item entry form provides the choices listed below to the person submitting an item with stimulus material, so these same designations should be used on any paper forms used when items are written. Using the following designations for types of stimulus materials on paper forms will facilitate later entry of the items on-line:

Type of Stimulus

- Written/Quote
- Music Notation
- Audio Sound Byte
- Arts Reproduction
- Photograph
- Video

Examples of Stimulus Materials:

Items that use stimulus materials often ask students to comprehend some important aspect of the stimulus, analyze statements, compare and contrast two short stimuli, evaluate statements made in the stimulus, or demonstrate some similar type of understanding through referring to the stimulus, reading the questions, and selecting the correct or best response. The following are some typical kinds of stimulus materials that can be used with multiple-choice items for the SCASS Arts Item Pool.

Written/Quote

Written, or textual, stimulus materials are probably the most common type of stimulus. Textual stimulus materials are fairly brief sections of text that is a quoted selection or written specifically for the item and used as part of an item. Obviously, any text that is quoted would need to be copyright-free. Examples of textual stimulus materials in the arts might be a section of a script or a few lines or paragraph(s) from a literary work in a theatre item, a section of a newspaper critique of a dance or music performance, a description of a print-making process in a visual arts item, or similar types of materials.

Music Notation, Arts Reproduction, or Photograph

Music notation, reproductions of art work, and photographs are often referred to as graphic stimulus materials. Graphic stimulus materials are generally still images. They can be simple marks, blocks of color, line drawings, halftones, color illustrations or photographs, charts, graphs, maps, diagrams, etc. Examples of graphic stimulus materials that might be used in arts items might be images of musical notation, a section of sheet music, reproductions of paintings or drawings, a set design sketch, a photograph of dancers demonstrating a particular concept, etc. Graphics used in items need to be copyright-free.

Audio

Audio stimulus materials are brief audio clips. The most familiar type of audio stimulus material is probably a recorded clip of a music performance. Other examples might include short excerpts of recorded spoken performances or other spoken information, such as oral critiques, oral directions, etc. Like quoted selections and graphics, audio clips also need to be copyright free.

Video

Video stimulus materials, of course, include motion and sound along with visual images and can include anything that can be captured on video. Obvious examples in the arts would be short video clips of theatre, dance, or music performance. Other examples might include video showing a sculpture from all sides or a mobile in motion, video showing performance art, video showing the responsibilities of someone in a particular career, such as a director of a play or an art museum curator or other arts career. Videos also need to be copyright-free.

Selection of Stimulus Materials

In addition to selecting stimulus materials that address Universal Design and Sensitivity issues, which are as much concerns with stimulus materials as with the wording and layout of items, there are four other major concerns that should be addressed when selecting stimulus materials to use with multiple-choice items:

1. copyright ownership,
2. relevance and appropriateness of the stimulus to the content and purpose of the item,
3. technical quality of the stimulus for clarity of interpretation by students and for printing or other types of reproduction for assessment purposes, and
4. technical specifications regarding specific types and sizes of files that the SCASS Arts Item Pool can accommodate are included on page 23.

Each of the four concerns mentioned above will be addressed below and continuing through page 23. Specific examples and discussion of stimulus materials that are problematic are presented in the constructed-response section of this training manual starting on page 24.

(1) Copyright Ownership

As noted earlier in this document, any stimulus used with an item must be free of copyright permission and ownership issues. As a rule, in any item development, if the copyright ownership of a stimulus or permission to use the stimulus has not been resolved prior to beginning to develop an item, then no matter how good the stimulus may be for that item, it should not be used. It can be a disappointing experience to spend time and energy developing an item with the hope that copyright permission can be obtained, only to discover later that permission is not granted or the fees for use are prohibitive.

For the SCASS Arts Item Pool, no copyrighted materials of any kind can be submitted to the on-line item review process. As noted earlier, stimulus materials to be used with items developed for the SCASS Arts Item Pool should be selected from the SCASS Arts pre-approved collections because there are no unresolved copyright issues with these stimulus materials. Other copyright-free collections may be considered on a case-by-case basis by the SCASS Arts Consortium.

(2) Relevance and appropriateness of the stimulus materials to the content and purpose of the item

All of the issues, concerns, and goals for developing high-quality multiple-choice items without stimulus materials that were addressed earlier, of course, also apply to developing multiple-choice items that use stimulus materials. Addressing some additional questions, such as those listed below, may help item developers to avoid some of the common pitfalls in using stimulus materials with multiple-choice items:

Is this stimulus really necessary?

It is often tempting to use stimulus materials because they are attractive elements in an item. However, if a stimulus isn't absolutely essential in order to ask a particular question, then a stimulus should not be included. When in doubt, leave the stimulus out.

From the user's point of view, stimulus materials take up room on a page and/or time from the overall test administration. Given the short amount of time students have to respond to multiple-choice items, unnecessary stimulus materials may slow students down and even cause confusion if they are trying to figure out why the stimulus was included.

Is this stimulus clear, succinct, and adequate for the question being asked?

Item developers should be sure that the content of stimulus selected is clear to students, especially since students will quite likely be encountering the stimulus for the first time and for a very short time. Whether the stimulus is a selection of quoted text, an art reproduction, audio sound byte, or some other type of stimulus, the content of that stimulus should clearly focus on what the stem of the item is asking and should not contain other distracting information that might cause confusion. The stimulus should be succinct but adequate to elicit appropriate responses from students.

Sometimes item developers are in a situation in which they have to "settle" for a stimulus that is close to what they need, but not quite what they would consider adequate. When the words in an item are not quite right, editing can often correct the problem, but if the "mis-fit" is in the stimulus itself, editing is usually not an option. Therefore, care must be taken not to "settle" for a stimulus that is not a good fit because it will likely result in an item that, when used, won't produce valid results.

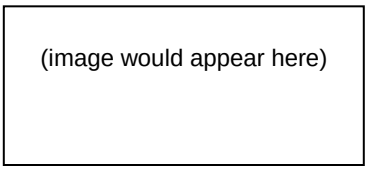
Does the placement/timing of the stimulus in the item communicate clearly?

This question relates to the formatting of the item. Depending on the stimulus, it may communicate best if students see/hear it before reading the stem, after reading the stem and before reading the options, or after reading the stem and the options. This decision will depend on the individual item. Whatever the format, the wording of the stem must correctly refer to the placement/timing of the stimulus and any information that is peculiar to the format of the item.

For example, if the stimulus is an image such as a reproduction of a work of art that would actually appear on the printed page the student is using, it is usually best to structure the item so that the student can read the item before or at the same time that he/she sees the image and in a normal left-to-right pattern, such as:

The painting shown below is an example of which artistic movement?

A. (Distractor)
B. (Distractor)
C. (Key)
D. (Distractor)

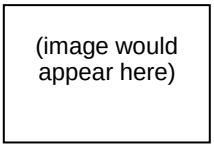


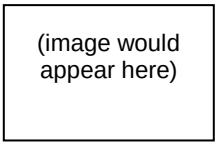
Note: For items that have more than one image as a stimulus or that have images as distractors and key, the entire item (including images) has to be submitted to the item review process as one graphic file. They should be formatted as follows:

Format Example #1

The paintings shown below are examples of which artistic movement?

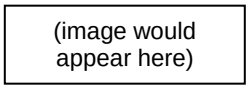
A. (Distractor)
B. (Distractor)
C. (Key)
D. (Distractor)

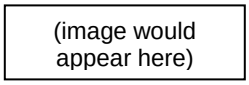


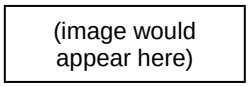


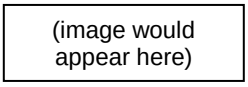
Format Example #2

Which painting is the best example of American Impressionism?

A. 

B. 

C. 

D. 

The example below should be used as a guide or model for developing items with audio or video clips, with the understanding that individual items may have requirements particular to those items. In general, such items for the

Item Pool should be written so the audio or video stimulus is repeated twice, and the clip should not exceed about 30 seconds. For example, if the stimulus is a 10-second audio clip of a music performance to show a particular concept, the item might be worded as follows:

> Read the question and the four answer choices below.
> A 10-second audio clip of a music performance will be played.
> The audio clip will be repeated so that you can hear it twice
> before you answer the question.
>
> Which of the following is an accurate critique of the music performance just played?
>
> A. (Distractor)
> B. (Key)
> C. (Distractor)
> D. (Distractor)
>
> *[Test administrator plays audio clip, pauses, repeats audio clip.]*

Note: Users of an item such as the one shown above may choose to modify the number of times the clip is repeated, if they feel a need to do so, but for consistency and reasonableness, items for the Item Pool should follow the model item structure above. It would be necessary for the users of such items to plan their test booklet to allow for playing the audio clip for all students at the same time. That is, if items with audio or video clips are included in a test with items that do not have audio or video clips, the items with audio/video clips probably would be grouped at the beginning of the test, end of the test, or a separate section of the test. Otherwise, students will get to these questions at varying times.

In any case, when developing multiple-choice items with stimulus materials, item developers should carefully consider not only the wording in the item and the content of the stimulus, but also how the words and stimulus might interact with each other. Developers should try to evaluate the item by playing the role of students who have never seen the item before, rather than viewing the item as they feel students in their own classrooms might, or to work with another item developer who has not seen the item before. A good rule to follow is to assume that there is some way that the item will be misinterpreted by a student, search for the possible interpretations that might happen, and then fix those "loopholes" if at all possible. The goal should be to create fair and unambiguous items that have as few as possible ways that a student might misinterpret them or interpret them in undesired ways.

As a note of caution, when using graphics, video, and audio as stimulus materials, because of familiarity with the image, the music, or the theatre or dance performance being depicted, it is often easy for item developers to assume that some concept or element can be clearly seen or heard, when in fact, to someone viewing or listening for the first time, it may not be clear. This familiarity with a work can cause us to mentally fill in gaps and cause us to assume that concepts are clearly represented in the audio or video even when they are not. Sometimes we have to force ourselves to look with fresh eyes and listen with fresh ears, as if we are students seeing or hearing a piece for the very first time. This is more difficult than it appears to be, but it is very important to the process of developing high quality items that are fair to students.

(3) Technical quality of the stimulus materials

The technical quality of stimulus materials is a very important consideration because the quality could directly affect the student's ability to answer the item correctly. Obviously, with still images, size, focus, contrast, color, and amount of fine detail, will all affect the ability of the student to comprehend the image clearly, so selecting a good quality image so that if the item is used on a test, the image will reproduce well enough for the student to be able to use it to answer the question. Images that have fewer fine lines and subtleties might be better choices. Line drawings, of course, will reproduce more easily and clearly than photographs or other halftones. Images that clearly focus, perhaps in close-up fashion rather than

from a distance, on whatever the item is asking about and contain no details that could conflict with or distract from the focus of the question are also important.

Sound quality and volume of both audio and video stimulus materials are critical to students' understanding of the item. Item developers should keep in mind competing ambient noise that may be present in a classroom during test administration. With a video stimulus, lighting conditions in the classroom during test administration can also be a factor that could interfere with students' ability to respond accurately to the item.

The length of a video or audio clip should be clearly specified in the item and be adequate to demonstrate the concept being asked about in the item, but short enough so as not to add unnecessary time to test administration. When considering length of time, item developers should keep in mind that students may need to hear/see the audio/video more than once.

In video selections, it is also important to consider the types of shots, camera distance from the subject (close-ups vs. wider views), and the angles of various shots. The question being asked in the item must match what is emphasized and clearly apparent in the video and vice versa.
Using multiple stimulus materials in an item can also cause unique problems that item developers need to consider. In items that use more than one stimulus, such as an item with images of two paintings that students are asked to compare or contrast, it is important to view the images together as the student will see them in order to make sure one doesn't affect the perception of the other or cause other unintended types of problems.

These are some of the most frequent issues item developers confront when using stimulus materials with items. Others that are unique to particular items will no doubt also occur. Using stimulus materials with multiple-choice and with constructed-response items expands possibilities, but it also introduces additional factors to be watchful for during the item development process.

(4) Technical Specifications for Stimulus Materials for the SCASS Arts Item Pool

The following table indicates the types of stimulus materials, file formats, and maximum file sizes that the Item Pool can currently accommodate:

SCASS Arts Item Pool—Specifications for Stimulus Materials

File format	Type of file	Max file size	Use-Dimension
jpg	Image	200KB	For photographs, color images. Approx. 600 pixels wide
gif	Image	200KB	For logos, line drawings. Approx. 600 pixels wide
doc	MS Word document	500KB	For print documents
ppt	MS Powerpoint	1MB	
pdf	Adobe portable document format	500KB	For print documents where it is imperative to keep formatting
wav	Audio file	1MB	
au	Audio file	1MB	
mp3	Audio file	1MB	
mid	Audio file	200KB	
ra/ram	Real Audio or Real Video	1MB	Single bit rate, not multi stream bits
mpg	Video	2-4MB	Use 1/2 dimension screen size (320x240 pixels) at approx. 20frames per second. 2-4 minutes long, as short as possible
avi	Video	2-4MB	Use 1/2 dimension screen size (320x240 pixels) at approx. 20frames per second. 2-4 minutes long, as short as possible
mov	Quicktime movie / video	2-4MB	Use 1/2 dimension screen size (320x240 pixels) at approx. 20frames per second. 2-4 minutes long, as short as possible

Section IV

DEVELOPING CONSTRUCTED-RESPONSE ITEMS, WITH OR WITHOUT STIMULUS MATERIALS

PURPOSE OF DEVELOPING CONSTRUCTED-RESPONSE ITEMS

In addition to including selected-response items (with and without stimulus materials), the SCASS Arts Consortium is moving forward to also start adding constructed-response items, with stimulus materials and without stimulus materials, to the SCASS Arts Item Pool. Constructed-response items allow for measuring students' understanding of certain Arts Standards that are difficult, if not impossible to measure with selected-response items. Constructed-response items also offer students a more individualized way of demonstrating their understanding of various arts concepts.

Constructed-response items include a variety of formats and are known by a wide variety of names, but they all share one thing in common. Constructed-response items are those in which students provide (construct) their own answers rather than selecting an answer from among the given options. Constructed-response items may range from those asking for brief factual answers that take less than a minute to respond to, such as "Who was the first president of the United States of America?" to much more elaborate items asking for extended responses in essay form or in the form of artistic creations and performances. Extended constructed-response items often ask students to analyze, evaluate, critique, develop, create, compare/contrast, justify, defend, etc. Extended response items, generally, are constructed-response items that allow the student to spend 5, 10, 30, 90 minutes, or more to respond and usually result in a variety of correct answers.

Regardless of whether the constructed-response item is brief and factual, extended, or somewhere in between, every constructed-response item also needs a set of appropriate criteria by which the responses should be scored. Scoring criteria (the "rules" by which responses to an item are scored) can be presented in a variety of ways: rubrics, checklists, narrative descriptions, scoring guides, etc.; but regardless of the format, the criteria should be as clear and as objective as they can be in order to remove as many subjective judgments from the scoring process as possible. For the SCASS Arts Item Pool, scoring criteria with score points 4 (highest), 3, 2, 1, and 0 are recommended, but there may be items for which fewer score points would be appropriate. Keep in mind that a multiple-choice item is the equivalent of a scoring scale of 1 and 0 (1=correct; 0=incorrect), so in order to make using a constructed-response item worthwhile, it should have more than a 1 and 0 scale. A scoring guide may contain additional descriptive information to help clarify the scoring criteria. An example of scoring criteria is included on page 39.

The best way to produce a high quality constructed-response item that can be scored fairly and consistently is to develop the item and the scoring criteria at the same time, try out the item with a sample of students, and then modify the item and criteria as necessary based on actual student responses. Scoring criteria are easier to understand and apply when they are accompanied by examples of student work representative of each score point. Because most of the constructed-response items for the item pool have not yet been tried out with students, examples of student responses are not available. A user of a constructed-response item is responsible for tryouts and field testing of these items and their scoring criteria and making revisions necessary.

Both the item and the scoring criteria need to be structured very carefully so as to elicit the intended kinds of responses from students and to be able to score those responses as fairly, objectively, and consistently as possible. To obtain valid and reliable measures of what students know and are able to do, as much ambiguity and subjectivity as possible must be removed from the item and its scoring criteria through careful development and tryouts with actual students.

As a general rule, because developing, answering, and scoring constructed-response items are time-consuming, expensive undertakings, constructed-response items should be used to measure concepts that are not possible to ask in selected-response form.

DEVELOPING CONSTRUCTED-RESPONSE ITEMS

The development of constructed-response items incorporates many of the same considerations that are part of the development of selected-response items as well as some additional considerations. As is the case in developing selected-response items for the SCASS Arts Item Pool, item developers are strongly encouraged to focus on developing constructed-response items that go beyond recall and basic knowledge items. Just because an item is a constructed-response item doesn't automatically mean that it measures some higher level of understanding. For example, the following item is a constructed-response item, but it is at a fairly low recall level, and it probably would be more efficient to ask this question in the form of a selected-response, multiple-choice item:

> Name an artist of the Italian Renaissance Period.
>
> _____

A common reaction to making this kind of item a higher level one might be to require more information, such as:

> Name three artists of the Italian Renaissance Period, and list one artwork produced by each of the three artists.
>
> _____
>
> _____
>
> _____

However, asking for more information of the same recall/knowledge level does not increase the thinking level required to answer the question. Therefore, in item development one should consider whether asking for more of the same kind of information really tells any more about what the student understands. If he/she can name two or three, does he/she understand any more about artists of this period, or does he/she just have a better memory?

Because constructed-response items are more time-consuming for students to respond to and more time-consuming to score than selected-response items are, it is prudent to be very selective about which kinds of items (selected-response or constructed-response) would be best to use to measure which kinds of student understanding.

Since the SCASS Arts item development began with selected-response items, the standards not readily compatible with the selected-response format have the fewest available items in all four disciplines. Although more selected-response items continue to be needed for the Item Pool, item developers might wish to focus first on developing constructed-response items for the Arts Standards where the fewest items currently exist, as per the "Item Needs Summary" for each discipline, pages 43-45.

CREATING A CONSTRUCTED-RESPONSE ITEM: A STEP-BY-STEP APPROACH

Getting Started:

The following approach is provided for the purpose of developing music, theatre, dance, and visual arts constructed-response items to add to and enhance, in a systematic way, the breadth and depth of the items in the SCASS Arts Item Pool.

*(Note: The rest of the material presented on this page is the same as that presented in the **Getting Started** portion of the selected-response item development section earlier in the training manual. If item developers have recently covered that information, the information below can be quickly reviewed before proceeding with the information on the next page.)*

High quality items are the foundation for high quality, valid and reliable assessments that measure students' understanding in a fair and unbiased manner. Meticulous attention during the item development process should produce items that are accessible to the widest range of students—sometimes referred to as "universal design." When developing items, item writers should avoid constructing items that might cause unnecessary challenges to students with disabilities or other special needs. Test items and stimulus materials should not demonstrate or inadvertently foster negative stereotypes or images. Some areas of concern are listed on the handout called ***Universal Design & Sensitivity Considerations*** as reminders, but item writers should use their best professional judgment regarding possible additional areas of concern. States that use items from the SCASS Arts Item Pool are responsible for sensitivity reviews and field testing items to detect possible bias. Item developers should refer to the handout throughout the item development and stimulus selection process.

Item developers should begin by reviewing the ***Item Needs Summary*** for the discipline for which they will be developing items. The Item Needs Summary lists the National Standards and grade levels of selected-response items already in the Item Pool as of May of 2003. For each discipline, a brief summary of needs for additional items is included.

After noting where the greatest needs are for the discipline for which they will be developing items, item developers should then choose a National Standard for which there are few, if any, items. All items for the Item Pool must be written to measure an important aspect of a National Standard. The ***National Standards*** handout included in this training manual lists a one-line descriptive statement for each standard in each discipline. For the SCASS Arts Item Pool, these one-line statements should be used to indicate which standard an item is designed to measure. To make the Item Pool useful for a variety of member states, there must be a common reference point. Therefore, the National Standards adapted statements listed will be used for that purpose. Ultimately, if a member state wishes to use items from the Item Pool, that state can match items to state standards or use whatever categorization system that is appropriate for the state's use of the items. Having the National Standards as a reference point will assist in that task.

Regarding grade level, item writers should concentrate on developing items for the grades with which they are most familiar, as items at all grade levels are needed. If an item developer is familiar with multiple grade levels, then he/she should concentrate first on developing items at grade levels where the greatest needs exist. However, in general, constructed-response items usually are not recommended or often used for kindergarten through grade 2 children, and use of constructed-response items for grade 3 children is fairly uncommon. Obviously, great care and expertise regarding the language development levels of children at these grade levels is required when developing items because the life experiences, vocabulary, and understanding of written words are still quite limited in very young children. If an item is suitable for more than one grade (e.g., grades 6-8), this grade range may be indicated. It is understood that users of items at the very early grades (i.e., K-2) at which students cannot yet read would most likely build their assessments so that the items would be read aloud to students.

DEVELOPING CONSTRUCTED-RESPONSE ITEMS

It is as challenging a task, if not more so, to craft a constructed-response item, as it is to create a high quality selected-response item. To meet the challenge of crafting high quality items for the SCASS Arts Item Pool, a recommended procedure that is similar to that recommended for developing multiple-choice items, with some differences as appropriate, is presented in this training manual.

The step-by-step item development procedure described in this document is based on procedures that have been followed for many years by many states and test development companies. It has been proven to work well when developing constructed-response items no matter what the subject area or discipline. States participating in the SCASS Arts Item Pool development process may choose to make some adaptations in the process to fit within the procedures of the member state; however, any adaptations should still result in the development of items that meet the needs and criteria of the SCASS Arts Item Pool.

As in developing multiple-choice items, it is also recommended that constructed-response item developers work in small groups or with a partner, rather than in isolation, even if they already have had extensive item development experience. Sharing of ideas and bouncing suggestions back and forth

among colleagues, at the item and scoring criteria creation level and when revising items and scoring criteria, usually results in higher quality items with clearer focus.

Constructed-response items are deceptively simple looking—just a few lines of direction with a space for students' responses. However, high quality constructed-response items are often more difficult to write than multiple-choice items, and, like the best multiple-choice items, are usually the product of many rounds of revision and tryouts. Item developers come to understand that revision, revision, and more revision is a large part of the normal process of constructed-response item development, just as it is with multiple-choice item development. It is important for new item developers to be aware that self-critiquing and peer-critiquing are normal and expected parts of developing high quality items. With constructed-response items in particular, it is often not until the item developers can see how students actually responded to the item that some of its short-comings may become apparent. Even the best-appearing constructed-response items can sometimes result in some unanticipated interpretations by students.

Item developers should each have a copy of the SCASS Arts Item Pool Style Guide included in this training manual to follow during item development. It is very important that all item developers follow this Style Guide so that all items will be in the same format, no matter which state develops the items. If the enclosed Style Guide varies from what the item writer may be accustomed to from other item development projects, the item writer must follow the SCASS Arts Item Pool Style Guide.

Keeping all of the previously stated information in mind, item developers should use the Worksheet for Developing a Constructed-Response Item included in this training manual for rough-drafting items (or create their own form containing the same information) and follow the step-by-step procedure below for constructing each item:

Note: For entry into the on-line Item Pool review process, each item must be given a short title to identify it. It may be easiest to name the item after it has been completed. The Item Title should be written on the Worksheet.

STEP 1: National Standard & Grade Level

Choose a National Standard and grade level upon which to focus, as described in the "**Getting Started**" section. Enter the Standard statement and grade level(s) that the item will address on the "Worksheet for Creating a Constructed-Response Item."

STEP 2: Information/Concept Focus of Item

Decide what important information or concept the item will address. Enter this information on the "Creating a Constructed-Response Item" worksheet.

What will this item attempt to find out from a student about this important information or concept? Be as clear and concise about what the item will measure as is possible at this point; it probably will be necessary, as the item progresses, to keep refining and honing its focus.

STEP 3: The Task

Briefly describe the task that the student will be asked to do. This information will help users of the Item Pool to understand quickly the gist of the constructed-response item.

Example: Students will watch two segments from the video series, Dancing, and will then write a short essay comparing the choreographic styles of George Balanchine and Twyla Tharp. Students will be allowed to make notes as they are watching the video. They will write their essays in their test booklets.

STEP 4: Materials/Setup

Not every item will require additional materials/setup, but if the item does have special requirements, ***briefly describe any materials, equipment, stimulus materials, or special setup necessary for the administration of the item.*** This information will help users plan appropriately if they choose to use the item.

Example: Enough paper for notes. A video cassette player and TV monitor. An edited video of the two segments from the series, Dancing, beginning with George Balanchine's "Meditations" and ending

with Twyla Tharp's "Sextet" rehearsal. The tape will be shown once to students. Students may make notes. The video will last 7 minutes; students will have 10 minutes to write their essays.

STEP 5: Directions & Script for Administrators

Many constructed-response items can be written to contain all the information students will need to respond appropriately. If an item is self-contained, and students can determine what they are expected to do from the stem and/or stimulus alone, then the item will not require directions and a script for test administrators. However, for more complicated constructed-response items, especially those that result in a performance assessment, an administrator's script is essential. The script is exactly what the person administering the assessment to students must say to the students so that all students receive uniform instructions.

Example of a partial script: "…after you have created your dance sequence, you will then perform the dance sequence. Your performance will be videotaped. …"

STEP 6: The Stem

Draft an item stem that addresses the concept or information identified in step two, keeping in mind that items measuring levels of thinking higher than recall/knowledge are needed most. Enter the draft item stem on the worksheet, directing the item toward an intended level of thinking (comprehension, application, interpretation, analysis, or evaluation).

The stem is usually a question, but the stem can be in the form of a statement to be completed. No matter which format the stem uses, it must be clear to students, from the stem alone (or stem and prompt, if used), what is being asked and how they are expected to respond.

Refine the wording of the stem to be as clear and concise as possible, paying attention to appropriate wording for the grade level. The natural tendency for adult item developers is to write at an adult level, so careful attention must be paid to appropriateness for students, especially at the elementary level.

Unlike with multiple-choice items that limit the length of the stem to no more than a line or two, constructed-response item stems can be as long as necessary to be clear. Usually a few lines or a brief paragraph is enough. Depending on the item, sometimes it is best to structure the item in "steps" to carry the student through. A variety of formats for constructed-response items can be used as necessary to elicit and capture the types of responses the item intends to get from students.

Finesse is required of the item developer in order to be specific enough without being too verbose. Usually a great deal of editing and rewriting is necessary to achieve that balance. A common tendency for item developers is to try to include too much in a constructed-response item. Item developers should be encouraged to focus on the important concept they hope the student understands, and construct the item to reveal that understanding. Including multiple concepts can be confusing to students and difficult to score fairly.

Similarly, another tendency is to ask students for multiple examples to show they understand a particular concept. Usually, one example is enough; multiple examples don't contribute additional information about students' understanding of a concept. Scoring responses with multiple-examples can be a complicated process, also.

STEP 7: The Scoring Criteria

Draft the tentative scoring criteria ("criteria" as used in this document means the rules/guidelines for scoring the expected kinds of responses students will most likely produce). Enter the tentative criteria on the worksheet. The criteria must describe, in as succinct a way as possible, the kinds of responses that would warrant the highest score (a typical score range for constructed-response items is 1-4, with 4 being the highest possible score) and the kinds of responses that would receive all other scores in the range (i.e., 1, 2, and 3), including scores of 0, and any other scoring anomalies, such as refusals, non-English responses, or whatever else might be pertinent to the particular item. For the Item Pool, items with score points of 0, 1, 2, 3, and 4 are preferred, although some items may warrant fewer score points.

Refine the wording of the criteria to be as clear and concise as possible, using language that is not ambiguous. Descriptions such as “nearly all”, “the majority of”, “very few”, etc., are too vague to use for scoring criteria that may be used by a wide variety of scorers.

Carefully consider the item and scoring criteria together. The criteria must cover what is asked for in the item and, at the same time, not give/subtract credit for things that are not asked for in the item. For example, if the item does not make it clear that writing an answer in essay form is expected, then the criteria cannot increase or decrease a student’s score based on whether or not the answer was in essay form. In such a case, the student who includes the correct information in a list would score the same as the student who included the correct information in essay form.

Taken together as a unit, the item and the criteria must address the intended information or concept at the item level intended. Often, writing the criteria will bring to light needed refinements in the item. This is not uncommon, and is a benefit of developing the item and the criteria at the same time. During the refinement and revising process of the item and the criteria, of course it is important not to lose sight of the intended purpose of the item and the Standard it purports to measure. It may be necessary to try different approaches in order to measure the desired concept.

STEP 8: Key Descriptor

For each item, ***select one “Key Descriptor” from the list*** on the Worksheet. The “Key Descriptors” help users search for particular kinds of items, such as those dealing with “Vocabulary” or “History/Culture” or some other type. If an item seems to fit in more than one of the given categories, item developers should select the one that seems to dominate. This information is only an aid in searching, so item writers should just use their best judgment as users of the Item Pool are likely to search in several categories to find items.

STEP 9: Level

Confirm the level of the item—knowledge/recall, or higher (comprehension, application, interpretation, analysis, or evaluation) and indicate the level on the worksheet. Sometimes, during the process of refinement, the item level is inadvertently or intentionally transformed from the level that was originally intended. A change in item level during the drafting process is okay when the change results in a better item. However, it is important to re-check the level and indicate the correct level on the “Developing a Constructed-Response Item” worksheet. Make the most reasonable judgment about its level that is possible, keeping in mind that the main distinctions we are concerned about are knowledge vs. higher level understanding. If it is clear that the item measures understanding at a level higher than the knowledge level, but which of the higher levels is being measured is difficult to distinguish, then item developers should use their best professional judgment as to the appropriate level.

STEP 10: Stimulus Materials

If the item incorporates stimulus materials, complete the Worksheet for **Indicating a Stimulus Used in an Item**, indicating a Stimulus Title, the Type of Stimulus, a Brief Description of the Stimulus, etc., as listed on the Worksheet.

STEP 11: Evaluate, Revise, and Finalize

Use the ***Checklist for Creating and Evaluating Constructed-Response Items*** included in this training manual to evaluate the rough-draft items regarding the significance of the item and the technical qualities of the item. The checklist is intended to assist in evaluating, revising, and finalizing items constructed for the Item Pool. Once all the boxes can be checked (indicating “okay”) then the item is ready to be presented to the item developer’s State Coach.

The section, “Explanation of the Checklist for Developing and Evaluating Constructed-Response Items,” which begins on the following page is intended to supplement and explain the criteria listed on the “Checklist for Developing and Evaluating Constructed-Response Items” form.

In addition to using the checklist and the explanation of the checklist during the drafting process, they should also be used to make sure an item is ready to present to the State Coach for possible entry into the Item Pool review process. (The Item Pool process is described in separate documentation.)

EXPLANATION OF THE CHECKLIST FOR CREATING AND EVALUATING CONSTRUCTED-RESPONSE ITEMS

The ***Checklist for Creating and Evaluating Constructed-Response Items*** included in this training manual should be used during the process of developing an item to guide its construction and after an item is developed to ensure that the item meets the criteria for consideration for the SCASS Arts Item Pool. Before submitting the item to the State Coach for review and possible entry into the Item Pool, the item developer(s) should check the item for significance and technical qualities using the checklist. If the item meets each criterion on the checklist, the item developer should check the box. If the item does not meet one or more of the criteria, the item should be revised until it meets all criteria. If it cannot be revised to meet all criteria, the item should not be submitted for review. If the checklist is used throughout the item development process, there should be very few items that will not meet the criteria at the final check stage.

The following information describes in more detail the criteria on the checklist. In some cases, an example of an item that is flawed and suggestions regarding how to revise the item are given.

Significance of the Item:

1. **The item relates directly to a National Standard.**

   Since the SCASS Arts Item Pool uses the National Standards as a common reference point for all member states, it is critical that each and every item developed/considered for the Item Pool relates clearly and directly to a National Standard. Item developers should use the one-line, general statements listed on the "National Standards for Item Entry Form" that indicate each of the National Standards. Item developers should not be concerned about pin-pointing exact subpoints given in the complete version of the National Standards, but rather they should focus on these general descriptive statements for item categorization. The general statement of the standard addressed will be indicated on the Item Pool entry form when the item is considered ready for entry. These statements will serve as a search guide to users of the Item Pool and will enable states to match items from the Item Pool to their own state standards, if they choose to do so. Constructed-response items should be developed for Standards which are difficult or impossible to measure with selected-response items. Each item should focus on only one Standard.

2. **The item deals with information or a concept of importance, not with trivia.**

   Using the selected National Standard as a guide should help item developers know what SCASS Arts considers important areas in each arts discipline. However, the item content must focus on important information/concepts within the National Standard. If one can answer the following question affirmatively, then the item is acceptable: If a student knows the answer to this question will that tell us something important about what the student knows, understands, etc.?

3. **The item is tightly focused so that it will be clear why a student gets it right or wrong.**

   A constructed-response item should try to measure a student's understanding of only one concept. It is easy to inadvertently introduce additional concepts, so a careful critique of what a student will have to know and do when responding completely and accurately is important. If an item is too ambitious or requires a student to demonstrate multiple concepts presented in a single item, then, it quite likely will be confusing to the student and result in responses that reflect that confusion, making them difficult to score consistently and fairly.

4. **If beyond recall or knowledge level, the item asks students to use higher-order thinking skills such as:**

   - Comprehending an important concept or information
   - Applying a concept/skill to a different situation
   - Making an interpretation/drawing a conclusion
   - Comparing/contrasting concepts or information
   - Analyzing and/or evaluating a claim or generalization

The emphasis for the SCASS Arts Item Pool is on developing items at levels higher than recall or knowledge. For purposes of this item pool, it is less important to tightly categorize items as to which of the higher levels an item is incorporating, than it is to distinguish between items at the recall/knowledge level vs. those at some higher level. Therefore, if it is difficult to draw fine lines between the higher levels, it is sufficient to use one's best professional judgment when classifying an item. For example, there might be analysis involved when comparing two concepts, or there might be application involved when evaluating a claim, etc. In these cases, the item developer should identify what he/she believes is the most appropriate classification level.

5. The item addresses sensitivity issues and the concept of "universal design" for the broadest audience.

A final check of the item for problems related to sensitivity issues that could produce a biased item is important. This would include a check to be sure the item has been written with the broadest possible audience of students in mind. The item should not cause unnecessary challenges to students with disabilities, English language learners, or others. The item should not present negative stereotypes toward any group, such as those groups listed on the "Universal Design & Sensitivity Considerations" handout. If there is any doubt that the item might be offensive to a particular group, favor one group over another, or exclude some students unnecessarily, the item should be reworked.

TECHNICAL QUALITIES OF THE ITEM AND SCORING CRITERIA:

1. The intent of the stem and what the student is expected to do is clear.

Make sure that each constructed-response item asks a clear question or otherwise provides clear and complete directions as to what is expected of the student. The stem should inform students exactly about what kind(s) of knowledge they are expected to demonstrate and exactly how they should demonstrate that knowledge. A simple method of testing a constructed-response item is for the item developer to try to respond to it as if he/she were the student.

Straightforward, objective wording is desirable. Using wording such as "comment on" or "give your observations of" is often vague and results in equally vague responses from students that are difficult to score. The more specific the directive, the better the chance that students will understand what is expected of them, and the clearer it will be from their responses whether or not they have understood the concept. Examples of direct statements include "explain why …" or "describe the effect x has on y" or "name the x and tell how it is used to y." Unless a student's opinion (rather than an answer) is truly sought, it is best to eliminate references to "you" and "your" in the stem. If the question asks the student, "Tell what you see in this set design that would tell you about the time period of the play," then almost any response has to be accepted because the question asks what the student sees rather than asking the student to "Tell one feature of this set design that indicates the time period of the play."

2. Information in the stem does not inadvertently cue the response.

Well-written constructed-response items should require students to use their own relevant knowledge and skills, rather than test-wiseness skills, to determine the correct answer. While it is true that sometimes students can use some fairly sophisticated analytical skills to find cues to an answer, the focus of the assessment should be on the students' comprehension of specific content and conceptual knowledge, and items that provide unintended cues for test-wise students should be avoided.

3. Negatives in the stem that would unnecessarily complicate understanding are avoided.

Stems using negative words or prefixes, such as not, un, non, least, which of these doesn't, or all of the following except, etc., generally serve to complicate an item. Difficult-to-read stems are more likely to measure students' ability to decipher the question than to measure their knowledge or skills. Usually, there are clearer ways to phrase the question. If the stem contains a negative, it should be reworked, if at all possible. In the process of reworking a negative stem, the criteria might have to be

reworded, also. In rare instances, it may not be possible to avoid the use of a negative in the stem. In these cases, extreme care is necessary in making the item clear to students.

4. **There is at least one plausible, accurate response to the stem.**

   The stem must be written in such a way that students can provide at least one plausible, accurate response. This may seem obvious, but it is not unusual to develop an item that appears sound, but if one actually attempts to respond to it, it becomes apparent that it does not work the way it was intended. Item developers should actually attempt to respond to the constructed-response item they create.

5. **The task demanded of the student is manageable within the timeframe.**

   Is the item asking for a reasonable response within the time limitations specified? The item should allow time for reading the question/directions and thinking about how to respond, as well as actually producing the response. A common mistake in developing constructed-response items is to ask students to do too much or to ask them to respond in a way that is more extensive or in a more complicated manner than is necessary for them to demonstrate that they understand the concept being asked about. Make sure the item is as streamlined and "user-friendly" as possible for the student. If a student does not score well, it should be because the student truly doesn't understand the concept being measured, not because he/she didn't understand what the item writer had in mind. What is expected of the student should never be a mystery.

6. **The item is in "user-friendly" format.**

   There are many ways items can be formatted, including with bulleted lists, lines that indicate to students how extensive their responses should be, an easy-to-follow format if the item asks for more than one response or a series of responses, adequate white space, etc., so that students don't have to spend time figuring out what they are to do. The item should carry the student through the task(s) in as clear a manner as possible.

7. **Language usage and grammar in the stem are correct.**

   It would be helpful to have an expert in language usage and grammar to review and edit items, especially if the item developers have questions about particular items they have written or if they feel less than confident about usage or grammar in particular items. Even if the content and concept of an item is excellent, a lapse in correct language usage can detract from all of the good qualities of the item.

8. **Superfluous wording in the item is avoided.**

   Keep the wording in the item to the point. Overwriting or inserting irrelevant information into the stem adds unnecessary or unfair complexity to the item. Avoid the temptation to include unnecessary instructional background information in the item. As difficult as it may be, avoid seizing the "teachable moment" when writing an item—just ask the question. It is important to keep the grade level of the students in mind, also. Using more difficult vocabulary does not raise the level of vigor of an item, it just makes it harder to read. For items that are intended to measure what students know and can do in the arts, reading levels that are too difficult may have a negative impact on the results. If a student doesn't do well, it will be difficult, and perhaps impossible to determine if the reason was because the student did not understand the artistic knowledge or concept the item was asking about or because of the reading comprehension level of the item.

9. **Complex formats in the stem and overall item are avoided.**

   Avoid the use of unnecessarily complex formats. Generally, it is a good idea to present items as clearly and simply as possible so that there is little doubt what the item is asking for and how the student should respond. There are occasions where multi-step formats can be useful to help guide the student through the item and provide specific spaces for responses to various steps. If an item uses a multi-step format, make every effort to keep the item as clear as possible. Remember that format complexity is not the same as item difficulty, and there is no good reason to use formats that will disadvantage some students who would otherwise have been able to respond adequately.

10. Scoring criteria for 4 (highest), 3, 2, 1, and 0 are unambiguous and complete.

Constructed-response items normally result in a variety of appropriate responses, a variety of partially acceptable responses, and a variety of responses that aren't accurate. While it is probably impossible to anticipate all the possible responses students might generate, the scoring guide should be written to cover as many as possible, including the full range of responses from best to worst, within the tentative scoring categories. Item writers often find it helpful to ask themselves, "what if" questions while they are developing constructed-response items and scoring criteria—"What if a student does X? What if a student doesn't do Y? What if a student does some of X, some of Y, and some of Z?" Terms that can be interpreted in different ways by different scorers should not be used (e.g., most, more than, usually, often, extensively, minimal, etc.) unless they are defined explicitly or quantified in some objective way. The goal of writing good scoring criteria is to eliminate as much ambiguity and subjectivity as possible.

11. The item and scoring criteria fit correctly/appropriately with each other, grammatically and conceptually.

After the item is considered "finished," it would be helpful to read the item and the scoring criteria one last time to make sure that not only all of the individual elements of the item are satisfactory, but also that all of the pieces hang together to make a whole item that is as coherent and smooth as possible. In fine-tuning the item and/or the scoring criteria, it is easy to lose sight of the whole item—the "can't see the forest for the trees" syndrome. It is especially important that students aren't asked for things that are not scored, nor scored for things that are not asked.

12. The item conforms to the SCASS Item Pool Style Guide for Constructed-Response Items.

To standardize the item format, each item must follow the Item Pool Style Guide for Constructed-Response Items.

13. Any stimulus materials fit with the stem, are of high quality, and are adequately identified.

Any prompt used with a constructed-response item must be clear, compatible, and consistent with what the item is asking about. If multiple stimulus materials are used, the appearance of the stimulus materials should be parallel—for example, several illustrations drawn in similar style and from a similar perspective. Stimulus materials must be of high quality and have been adequately identified for entry into the Item Pool.

14. Any stimulus materials conform to the technical specifications of the Item Pool.

Stimulus materials must meet the technical specifications of the various file types described on page 23.

15. A brief description of the gist of the task is provided.

For each constructed-response item, a brief description of the task(s) students are expected to complete is provided.

16. Any special materials, equipment, or setups have been adequately described.

For constructed-response items that require any special materials such as art supplies, equipment such as a VCR, or special setup of the room, etc., have been described for a potential user of the item.

17. Directions and script for assessment administrators have been provided (for other than short constructed-response items).

For more complex items, especially performance type constructed-response items, complete directions and a script for assessment administrators to read to students have been provided.

EXAMPLE CONSTRUCTED-RESPONSE ITEMS AND STIMULUS MATERIALS

The following are examples of draft constructed-response items that incorporate stimulus materials. Each item, and its stimulus material are critiqued, and discussion-starter questions are included on the following pages to provide item developers with some guidance as to what kinds of problems the draft items might contain and how they might be improved. These example items may be used as handouts for training purposes.

Example #1: Draft Constructed-Response Items with Stimulus

In this example, two items relate to one stimulus which is a photograph (see below). Critiques of the two items, the stimulus, and the additional item information appear on the next two pages.

1. By looking at the photo below, identify three things you can project about this play with regard to the time period, presentational style, theatrical genre, etc.? Support your statements with observations from the photo. (5 minutes)

_____

[Note: additional lines provided for student response, but not reproduced here]

Using this same prompt, comment on which actor has the focus in this picture and why? (5 minutes)

_____

[Note: additional lines provided for student response, but not reproduced here]

National Standard: 3. Designing by conceptualizing & producing artistic environments

Verifiable Source: Scene Design and Stage Lighting, 6th Edition, W. Oren Parker & R. Craig Wolf, Harcourt Brace Jovanovich College Publishers, New York, ISDN 0-03-028777-4

Critique/Discussion of Example #1

About the first item in Example #1—

The National Theatre Standard that this item is attempting to measure is, ***"Standard: 3. Designing by conceptualizing & producing artistic environments."*** The item stem by itself does appear to ask a question appropriate for this standard. However, the stimulus may not adequately depict information about the design of the artistic environment for a student to be able to answer the question completely. Information about the time period can be detected from the costumes, but is there enough information in the stimulus for a student to accurately conclude anything definitively about the theatrical genre or presentational style? What does "presentational style" mean to a student? Because "presentational style" is a phrase that is commonly applied in a variety of ways, it is probably not precise enough to communicate the same meaning to all students.

The wording of the stem could be unclear to some students. Is the student expected to identify three things about the time period, three things regarding presentational style, and three things regarding theatrical genre; or is the student expected to identify a total of three things: one regarding time period, one regarding presentational style, and one regarding theatrical genre; or is it okay if the student identifies three things regarding time period? How is the student expected to respond to the "etc."? Does this imply that there might be extra credit given if the student thinks of other things that can be concluded from the picture?

Is the phrase "things you can project about" clear? Is there a more direct way of stating what the student is expected to do? How are students expected to "support [their] statements with observations from the photo"? For example, if a student states that the costumes indicate the correct time period, what "observations from the photo" could be used to support that statement? Given that the student has only 5 minutes to read the item, analyze the stimulus, and respond with three statements, each with a piece of supporting evidence, is this item trying to do too much? If the student provides 3 things, are we any more certain that he/she understands the theatre concept than if he/she provided 2 things or one?

This item could be improved by editing out the references to "you" and "your" to make the item more objective. Then, the wording could be streamlined and made more precise. Finally, a bit of reformatting might help. There are a variety of ways that this item could be restated, depending on the stimulus that is selected, but one way to improve the item might be as follows:

> Use the picture below to answer these three questions.
>
> What is the time period of the play?
>
> What is the theatrical genre of the play?
>
> What information in the picture helps communicate the time period and the theatrical genre?
>
> _____
>
> ***[with additional lines provided]***

Another approach might be to ask the student to simply:

> Describe how information in the picture below helps communicate the time period and theatrical genre of the play?
>
> _____
>
> ***[with additional lines provided]***

About the stimulus for the first item in Example #1—

The stimulus seems to have some misfit problems with the item stem that would require the selection of a different photograph. This situation might be improved: (1) by selecting a photo that would more

clearly give information about the time period and theatrical genre, (2) by editing the stem, and (3) by the addition of some formatting to clarify what students are expected to do.

About the second item in Example #1—

Again, the National Theatre Standard that this item is attempting to measure is, "Standard: 3. Designing by conceptualizing & producing artistic environments." As with the first item, this item by itself appears to be measuring this Standard; the presumption is that there is something about the designed artistic environment in the stimulus that can be used to answer the question.

However, the stimulus, unfortunately does not communicate through lighting design, set design, or costume design adequately to determine which actor has the focus in the picture. Rather it is the position of the actors relative to each other that communicates focus. The actor on the right is giving focus to the actor on the left. Therefore, even though the stem is asking a design-related question, the stimulus is not clearly providing a design-related answer to the question. There is a misfit between stem and the stimulus, which probably can only be corrected by replacing the stimulus.

The stem is concise and objective, but as a general rule, phrases such as "comment on" are usually too vague for students to interpret as the item writer desires. Regarding format for the Item Pool, each item has to include the stimulus, rather than having two separate items referring to a stimulus as those using the Item Pool may wish to select one item, but not the other. Therefore, they need to function as independent items. One way to improve this item might be:

In the picture below, which actor has the focus and why?

_____

[with additional lines provided]

About the stimulus for the second item in Example #1—

As noted in the discussion above regarding the first and second items, the stimulus seems to have some misfit problems with the item stems that would require the selection of a different photograph. Some other considerations for selecting a replacement stimulus to use with these items might be: a photo with better overall composition that is more artistically interesting and indicative of good stage direction, a photo that depicts better lighting design instead of having the center spotlight with the actors to the right and left, a closer-up shot that would reveal facial expressions, etc.

About the additional item information for Example #1—

Information required for the Item Pool is listed on the "Worksheet for Creating a Constructed-Response Item" and the "Worksheet for Indicating a Stimulus Used in an Item." This Example Item includes, most, but not all of the required information. For instance, the "Key Descriptor" is not indicated for this example item. Before an item is ready to be submitted to the on-line Item Pool for review, all of the information required for item entry into the Item Pool must be provided.

National Standard: 3. Designing by conceptualizing & producing artistic environments

Verifiable Source: Scene Design and Stage Lighting, 6th Edition, W. Oren Parker & R. Craig Wolf, Harcourt Brace Jovanovich College Publishers, New York, ISDN 0-03-028777-4

Example #2: Draft Constructed-Response Items with Stimulus

In this example, the item refers to a stimulus that is a set design sketch (see below).

A critique/discussion of the item, the stimulus, and the additional item information appears on the next page.

1. By analyzing the set design above, what inferences may be drawn about this play and the characters portrayed in this scene? What do you see in this drawing that leads you to these conclusions? Consider social status, theatrical genre, historical era, etc.

   _____
   _____
   _____
   _____
   _____
   _____
   _____

   [with additional lines provided]

National Standard: 3. Designing by conceptualizing & producing artistic environments

Verifiable Source: Scene Design and Stage Lighting, 6th Edition, W. Oren Parker & R. Craig Wolf, Harcourt Brace Jovanovich College Publishers, New York, ISDN 0-03-028777-4

Critique/Discussion of Example #2

About the item in Example #2—

The National Theatre Standard that this item is attempting to measure is, "STANDARD: 3. DESIGNING BY CONCEPTUALIZING & PRODUCING ARTISTIC ENVIRONMENTS." The item stem appears to address this standard, as does the stimulus. Aside from being somewhat ambiguous in what it is asking students to do, the item appears to be an adequate item. The questions of whether the stimulus provides adequate information to answer all of the questions posed in the stem and whether or not the stem is overly ambitious in what it is asking of students are similar to those questions asked of Example #1. Is there enough depicted in the sketch to indicate theatrical genre, for instance? If not, perhaps the item should ask about social status and historical era only.

As with Example #1, the stem in Example #2 could be streamlined and reformatted to improve clarity. One way to improve the item might be:

1. Based on analysis of the set design above, describe the conclusions that can be drawn regarding:

   a. the social status of the characters portrayed in this scene, and
   b. the historical era of this play.

_____
_____
_____
_____
_____

[with additional lines provided]

About the stimulus in Example #2—

Assuming that the stimulus and the item stem fit together adequately, as noted above, the stimulus appears to be adequate. However, since the item is asking students to base their responses on analysis of the set design, the notations "for Puttin' on the Ritz '98" and the title "The Glass Menagerie" should be deleted as students may answer based on their knowledge of the play, rather than from their analysis of the set design.

About the additional item information for Example #2—

Information required for the Item Pool is listed on the "Worksheet for Creating a Constructed-Response Item" and the "Worksheet for Indicating a Stimulus Used in an Item." As was the case with Example #1, Example#2 also includes, most, but not all of the required information. Individual states may need to record additional information for their own purposes, but this information is not entered into the SCASS Arts Item Pool. Before an item is ready to be submitted to the on-line Item Pool for review, all of the information required for item entry into the Item Pool must be provided.

Suggested Improvements—

On the following page, a revised version of this item is presented along with a proposed set of scoring criteria. The revision would crop off the verbal identification at the bottom of the sketch, place the stem before instead of after the stimulus, and the wording of the stem would be modified to communicate more clearly and directly the task to be done in order to adequately respond to this item.

Example #2: Suggested Revision (may still require additional revision)

The revised version of the item would use the same sketch as a stimulus, but the stem would be revised as follows:

> By looking at the set design below, you can draw inferences about the play. In essay form, describe those inferences, including in your answer the type of set, era, social class, and mood. Support your answers with specific observations from the drawing.
>
> [set design sketch, with wording cropped off, would appear hear]
>
> [sufficient lines for student essay response would follow the sketch]

Scoring Criteria: Acceptable responses for each of the four aspects that must be addressed in the response:

Type of set—As long as the answer is supported accurately, the response could be either a box set, proscenium set, realistic set, big set, multilevel set; possible references include: 3 walls, defined arch

Era—As long as the answer is supported accurately, the response could be: Victorian, 1800s, early 1900s, or 1700s; possible references include style of dress, style of furniture, use of candles, ceiling height

Social Class—As long as the answer is supported accurately, the response could be upper class, upper/middle class, or middle class; possible references include: furniture, line, chandelier, windows, drapery, linens, and the look/appearance of the woman

Mood—As long as the answer is supported accurately, the response could be any of the following: light, upbeat, happy, cheerful, lonely, isolated, secluded, mellow, or calm; possible references: color choices, fabric, greenery, lights

(Note: For ease of use, the above descriptions are not repeated in the score point descriptions below.)

Score Description

4	The response clearly, coherently, and insightfully analyzes the scenic designer's rendering, supporting the inferences with specific observations from the drawing. The response accurately addresses and supports inferences regarding all four of the aspects as described above.
3	The response clearly addresses the majority (3 out of 4) of the aspects called for in the question, with accurate support. However, the response fails to address all four of the categories or fails to support some of the observations made.
2	The response addresses only 2 of the aspects called for and provides accurate support for the observations made. The response may or may not attempt unsuccessfully to address and/or support other observations.
1	The response accurately addresses only 1 aspect of the question and may or may not provide accurate support for that observation. The response is an obvious attempt to answer the question, but it has significant inaccurate information, or it is incomplete, unclear, or confused.
0	The response does not accurately address any of the 4 aspects, and does not provide accurate support. A "0" response may be completely inaccurate, indecipherable, a refusal to respond, a comment such as "I don't know the answer", or a blank paper (no response).

Developing Strong and Fair Scoring Criteria:

Strong and fair scoring criteria should direct the scorer, as clearly as possible, to which aspects of a response should be scored and which should not. What is scored should match what is asked—no more, no less.

A well-constructed item, developed along with carefully-created scoring criteria and accompanied by adequate training of scorers can yield scores that are reliable to a fairly high degree. It is highly recommended that training include pre-selected responses that represent each score point. These are often called “anchor papers” because they serve as reference points by which scorers anchor their interpretations of the score point descriptions in the scoring criteria. Training sessions with opportunities to practice scoring responses, discuss scores assigned and ask questions and clear up any remaining misperceptions before scoring “live” papers are also recommended.

The purpose of developing scoring criteria and providing training is to eliminate subjectivity and increase the reliability of scoring, to the extent possible. Ultimately, the student and the scorer should both feel confident that the score assigned to a response accurately describes the response. There are at least three common quality control measures that are used in scoring responses to constructed-response items:

Inter-rater Agreement—This quality control measure checks consistency of scores between and among scorers. It involves at least two scorers (who have been adequately trained) scoring each response independently and comparing agreement rates. Typically, an agreement rate of around 80 per cent or higher is expected and can be achieved. Agreement rates in the upper 80 per cents and even over 90 per cent are possible. For any response on which two independent scorers disagree, a third scorer usually resolves the difference.

Intra-rater Agreement—This quality control measure checks the consistency of scores assigned by individual scorers—i.e., a scorer’s internal agreement. The purpose of this internal consistency check is to make sure that the score is not changing his/her internal criteria over time. The most common method of checking for this internal consistency is to periodically insert pre-scored responses into the responses to be scored, without the scorer’s being aware, of course. Then, a check is made to see if the score assigned matches the score already assigned. Intra-rater agreement rates are usually expected to be 80 per cent or higher.

Validity—This quality control measure checks to make sure scores are consistent with the scoring criteria. Even if there are high inter-rater and intra-rater agreement rates, unless there is also a high degree of consistency of those scores with the scoring criteria, the validity of the scores could be called into question. If scores are not valid, it means that they are not measuring what they purport to be measuring. Typically, validity checks are made by an expert scorer who does random checks of scores assigned to make sure that the scores are in agreement with the criteria.

All of these factors work together to produce valid, reliable measures of students’ understanding, as demonstrated through constructed-response items. The item, stimulus (if used), and the scoring criteria should be considered as equally important components of a strong, fair constructed-response item. If these components are all sound, and the training of scorers is adequate, high-quality results can be expected—valid scores with high inter-rater and intra-rater agreement rates. When care is taken with all of the components, it is possible to create constructed-response measures that are highly reliable.

Section V

HANDOUTS, WORKSHEETS, CHECKLISTS, & FORMS

The following section of this training manual contains reproducible handouts and worksheets that can be used to capture and to evaluate items on paper before they are entered into the Item Pool's on-line review process. They are designed to record all the essential information necessary for on-line entry, but states may adapt them to suit additional state needs.

Handouts include the following:

Universal Design & Sensitivity Considerations

Item Needs Summary (as of May 2003)—Dance
Item Needs Summary (as of May 2003)—Music
Item Needs Summary (as of May 2003)—Theatre
Item Needs Summary (as of May 2003)—Visual Art

National Standards

Style Guide for Multiple-Choice Selected-Response Items
Style Guide for Constructed-Response Items

Worksheets/Checklists include the following:

Worksheet for Creating Multiple-Choice Items
Worksheet for Creating Constructed-Response Items
Worksheet for Indicating a Stimulus Used in an Item

Checklist for Creating and Evaluating Multiple-Choice Selected-Response Items
Checklist for Creating and Evaluating Constructed-Response Items

Forms include the following:

Assignment of Copyright
Release Form

UNIVERSAL DESIGN & SENSITIVITY CONSIDERATIONS

FOR ITEM DEVELOPMENT AND STIMULUS SELECTION

High quality items are the foundation for high quality, valid and reliable assessments that measure students' understanding in a fair and unbiased manner. Meticulous attention during the item development process should produce items that are accessible to the widest range of students—sometimes referred to as "universal design." When developing items, item writers should avoid constructing items that might cause unnecessary challenges to students with disabilities or other special needs. Test items and stimulus materials should not demonstrate or inadvertently foster negative stereotypes or images. Some areas of concern are listed below as reminders, but item writers should use their best professional judgment regarding possible additional areas of concern. States that use items from the SCASS Arts Item Pool are responsible for sensitivity reviews and field testing items to detect possible bias.

Age

Age Appropriateness

Controversial Points of View or Subject Matter

Cultural or Language Differences

Disability

Economic Advantage/Disadvantage

Ethnicity

Gender

Gender Orientation

Geographic/Regional

Magic/occult

Religion

Rural/Urban

Violence or Other Inappropriate Behaviors (stealing, cheating, etc.)

Other

ITEM NEEDS SUMMARY FOR DANCE

(AS OF MAY 2003)

The following tables summarize information regarding the selected-response, multiple-choice dance items that were in the SCASS Arts Item Pool as of May 6, 2003.

As shown in the tables below, there are items at all grade levels, with the greatest need for items at the elementary grade levels. While more items measuring all standards in dance are needed, the greatest need for items is for standards 2, 3, 4, 6, and 7. As a reminder, many items are suitable for use at several grades, so they are counted at every grade at which they can be used; the total of 394 in the table showing items by Grade does not mean that there are 394 unique items for dance in the Item Pool.

DANCE	
Grade	**No. of Items**
K	**9**
1	**9**
2	**14**
3	**17**
4	**23**
5	**27**
6	**36**
7	**41**
8	**44**
9	**41**
10	**45**
11	**44**
12	**44**
Total	**394**

National Dance Standards Addressed	
Standard #	**No. of Items**
Standard #1	**21**
Standard #2	**11**
Standard #3	**9**
Standard #4	**0**
Standard #5	**33**
Standard #6	**1**
Standard #7	**11**
Total	**86**

ITEM NEEDS SUMMARY FOR MUSIC

(AS OF MAY 2003)

The following tables summarize information regarding the selected-response, multiple-choice music items that were in the SCASS Arts Item Pool as of May 6, 2003.

As shown in the tables below, most of the items are at the middle and high school levels, with no items at kindergarten or grade 1 and only 2 at grade 2. Therefore, the greatest need currently is at the elementary level. While more items measuring all standards in music are needed, the greatest need for items is for standards 1, 2, 3, and 4 and for standards 7 and 8. As a reminder, many items are suitable for use at several grades, so they are counted at every grade at which they can be used; the total of 406 in the table showing items by Grade does not mean that there are 406 unique items for music in the Item Pool.

MUSIC	
Grade	**No. of Items**
K	**0**
1	**0**
2	**2**
3	**17**
4	**29**
5	**40**
6	**60**
7	**60**
8	**59**
9	**45**
10	**36**
11	**29**
12	**29**
Total	**406**

National Music Standards Addressed	
Standard #	**No. of Items**
Standard #1	**2**
Standard #2	**2**
Standard #3	**2**
Standard #4	**6**
Standard #5	**14**
Standard #6	**32**
Standard #7	**4**
Standard #8	**6**
Standard #9	**29**
Total	**97**

ITEM NEEDS SUMMARY FOR THEATRE

(AS OF MAY 2003)

The following tables summarize information regarding the selected-response, multiple-choice theatre items that were in the SCASS Arts Item Pool as of May 6, 2003.

As shown in the tables below, most of the items are at the high school level, with few items at kindergarten through grade 4. Grades 5 and 6 have only 14 and 18 items, respectively. Therefore, the greatest need currently is at the elementary level, with a need also at the middle school level. While more items measuring all standards in theatre are needed, the greatest need for items is for standards 1, 3, 4, 6, 7, and 8. As a reminder, many items are suitable for use at several grades, so they are counted at every grade at which they can be used; the total of 328 in the table showing items by Grade does not mean that there are 328 unique items for theatre in the Item Pool.

THEATRE	
Grade	**No. of Items**
K	**3**
1	**3**
2	**3**
3	**4**
4	**4**
5	**14**
6	**18**
7	**20**
8	**26**
9	**58**
10	**59**
11	**58**
12	**58**
Total	**328**

National Theatre Standards Addressed	
Standard #	**No. of Items**
Standard #1	**5**
Standard #2	**18**
Standard #3	**9**
Standard #4	**9**
Standard #5	**20**
Standard #6	**4**
Standard #7	**2**
Standard #8	**4**
Total	**71**

ITEM NEEDS SUMMARY FOR VISUAL ARTS

(AS OF MAY 2003)

The following tables summarize information regarding the selected-response, multiple-choice Visual Arts items that were in the SCASS Arts Item Pool as of May 6, 2003.

As shown in the tables below, most of the items are at Grade 3 and above, with the bulk of the items at the middle and high school levels. Therefore, the greatest need currently is at the elementary level, especially at grades K through 2. While more items measuring all standards in theatre are needed, the greatest need for items is for standards 3, 5, and 6. As a reminder, many items are suitable for use at several grades, so they are counted at every grade at which they can be used; the total of 587 in the table showing items by Grade does not mean that there are 587 unique items for visual arts in the Item Pool.

VISUAL ARTS	
Grade	**No. of Items**
K	**1**
1	**2**
2	**7**
3	**25**
4	**33**
5	**60**
6	**68**
7	**79**
8	**79**
9	**72**
10	**62**
11	**49**
12	**50**
Total	**587**

National Visual Arts Standards Addressed	
Standard #	**No. of Items**
Standard #1	**46**
Standard #2	**29**
Standard #3	**7**
Standard #4	**29**
Standard #5	**4**
Standard #6	**11**
Total	**126**

NATIONAL STANDARDS FOR ARTS

DANCE

1. Identifying and demonstrating movement elements and skills in performing dance
2. Understanding choreographic principles, processes, and structures
3. Understanding dance as a way to create and communicate meaning
4. Applying and demonstrating critical and creative thinking skills in dance
5. Demonstrating and understanding dance in various cultures and historical periods
6. Making connections between dance and healthful living
7. Making connections between dance and other disciplines

MUSIC

1. Singing, alone and with others, a varied repertoire of music
2. Performing on instruments, alone and with others, a varied repertoire of music
3. Improvising melodies, variations, and accompaniments
4. Composing and arranging music within specified guidelines
5. Reading and notating music
6. Listening to, analyzing, and describing music
7. Evaluating music and musical performances
8. Understanding relationships between music, the other arts, and disciplines outside the arts.
9. Understanding music in relation to history and culture.

THEATRE

(PARAPHRASED FOR THE ITEM POOL)

1. Script writing by improvising, planning, writing, and refining
2. Acting by role playing, improvising, and portraying characters
3. Designing by conceptualizing and producing artistic environments
4. Directing by planning, organizing, and conducting rehearsals and dramatizations
5. Researching by finding, using, evaluating, and synthesizing cultural and historical information to support artistic choices
6. Comparing and integrating theatre with other art forms
7. Critiquing informal and formal theatre, film, television, and electronic media productions
8. Understanding context by analyzing the role of theatre, film, television, and electronic media

VISUAL ARTS

1. Understanding and applying media, techniques, and processes
2. Using knowledge of structures and functions
3. Choosing and evaluating a range of subject matter, symbols, and ideas
4. Understanding the visual arts in relation to history and cultures
5. Reflecting upon and assessing the characteristics and merits of their work and the work of others
6. Making connections between visual arts and other disciplines

STYLE GUIDE FOR MULTIPLE-CHOICE SELECTED-RESPONSE ITEMS

1. State the stem in question form if possible. For example:

   What type of design uses shapes and forms without reference to reality?

   Rather than-

   A design that uses shapes and forms without reference to reality is

2. If you do not use the question form, do not put a colon at the end of the stem.
3. If you do not use the question form, put periods at the end of each option (see example below).
4. There should be four options; one key and three distractors labeled A, B, C, and D with each letter followed by a period, as shown in the examples below.
5. Excluding proper nouns and titles, do not use initial caps in the test item key or distractors, for example:

   A. a visual and performing arts council
   B. professional dance groups
   C. Mozart and Shakespeare
   D. President's Advisory Committee on the Arts

General Guidelines

1. Write the key (the key is the only correct answer) after the item (see examples below for format).
2. Spell out ages and grades, for example: 25 thirteen-year-olds in eighth grade
3. Spell out centuries and decades, for example: twentieth century
4. Create items in 12 point Times New Roman font.

Examples of Multiple-Choice Items Without Stimulus Materials

1. Which instrument uses the bass clef for its notation? A. saxophone B. clarinet C. trombone D. trumpet Key: C. trombone	2. An actor's speech heard by the audience but supposedly not by other characters on stage is known as an A. apology. B. aside. C. undertone. D. upstage action. Key: B. aside.

(see next page for examples of multiple-choice items with stimulus material)

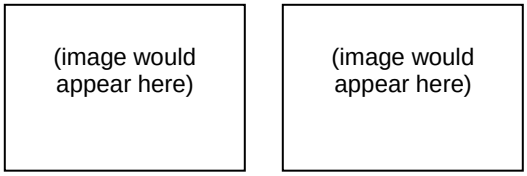
STYLE GUIDE FOR MULTIPLE-CHOICE SELECTED-RESPONSE ITEMS *CONTINUED*

EXAMPLES OF MULTIPLE-CHOICE ITEMS WITH STIMULUS MATERIALS

EXAMPLE (format of constructed-response item with two graphic images as stimuli; in this case, the entire item would need to be submitted as a pdf file):

The two paintings shown below are examples of which artistic movement?

A. (Distractor)
B. (Distractor)
C. (Key)
D. (Distractor)



EXAMPLE (format of constructed-response item with audio clip as a stimulus):

Read the question and the four answer choices below.

A 10-second audio clip of a music performance will be played. The audio clip will be repeated so that you can hear it twice before you answer the question.

Which of the following is an accurate critique of the music performance just played?

A. (Distractor)
B. (Key)
C. (Distractor)
D. (Distractor)

[Test administrator plays audio clip, pauses, repeats audio clip.]

STYLE GUIDE FOR CONSTRUCTED-RESPONSE ITEMS

GENERAL GUIDELINES

Because there is more variety among formats of constructed-response items, the guidelines are less prescriptive than for selected-response items.

1. Spell out ages and grades, for example: 25 thirteen-year-olds in eighth grade
2. Spell out centuries and decades, for example: twentieth century
3. Create items in 12 point Times New Roman font.
4. Use adequate margins, white space, etc., to make the item easier to read.
5. Formatting, such as breaking up the tasks within the item, presenting bulleted lists of important points students are required to address, and layouts that takes students through the steps/parts of the item are recommended.
6. Provide space, lines, boxes, or other appropriate area for the student response. Make sure the amount of space provided is compatible for the response. For example, if the task is to name something, then a single line is probably adequate; if the task is to write a 10-minute essay, at least a blank half-page or a half page of lines would be appropriate.
7. Do not use boldface, italics, or underlining, as the on-line item entry process does not accommodate these special editing/type features. Quotation marks should be used for titles mentioned within an item, if needed.
8. To the extent possible, use a consistent style for constructed-response items and scoring criteria, based on the examples in the Training Manual.
9. If in doubt, consult with the SCASS Arts representative.

WORKSHEET FOR CREATING A MULTIPLE-CHOICE SELECTED-RESPONSE ITEM

USE THIS WORKSHEET FOR DRAFTING, REVIEWING, EVALUATING, REVISING, AND FINALIZING MULTIPLE-CHOICE ITEMS.

Item Title _____

(a short name to identify this item as distinct from others; can be done after item is complete)

1. **National Standard**—What national standard are you assessing with this item? (Enter the one-line statement from the "National Standards—Adapted for the Item Entry Form")

   At which grade level(s)?

   _____ Grade(s) ___

2. What are you **attempting to find out** with this item/what is its **focus**/why **ask this question**/why is it **important**?

   _____

   _____

3. Write the **stem** (question) here:

   _____

   _____

4. Write the **key** (correct answer) here:

   _____

5. Write the three plausible **distractors** here:

   (1) _____

   (2) _____

   (3) _____

6. Choose only <u>one</u> of these "**Key Descriptors**" to indicate one that could be used to search the Pool for this item:

   ❒ Vocabulary

   ❒ Composition/Improvisation/Performance

   ❒ History/Culture

   ❒ Analysis/Evaluation

   ❒ Careers/Interdisciplinary Connections

7. Check which **level of understanding** the item is assessing.

   ❒ Recall/Knowledge ❒ Higher (Comprehension, Application, Interpretation, Analysis, or Evaluation)

8. If this item incorporates stimulus material, complete the "Worksheet for Indicating a Stimulus Used in an Item;" the stimulus must be submitted along with the item. If no stimulus material, use the "Criteria for Creating and Evaluating Multiple-Choice Items" checklist to review, evaluate, revise, and finalize the item drafted above.

WORKSHEET FOR CREATING A CONSTRUCTED-RESPONSE ITEM

USE THIS WORKSHEET FOR DRAFTING, REVIEWING, EVALUATING, REVISING, AND FINALIZING CONSTRUCTED-RESPONSE ITEMS.

Item Title _____
(a short name to identify this item as distinct from others; can be done after item is complete)

1. **National Standard**—What national standard are you assessing with this item? (Enter the one-line statement from the "National Standards—Adapted for the Item Entry Form") At which grade level(s)?

_____ Grade(s) ___

2. What are you **attempting to find out** with this item/what is its **focus**/why **ask this question**/why is it **important**?

_____

_____

3. Write a brief description of the task (what item expects students to do):

_____

_____

4. Describe any **materials or equipment needed** (i.e., VCR, audio equipment, etc.) and special set up required for administering the item, if any (attach additional page(s), if necessary).

_____

5. Attach **directions and script** for assessment administrators, if needed.

6. Write the **stem** here; attach additional page(s) and/or format layout if necessary:

_____

_____

7. For each dimension being scored, attach scoring criteria indicating a description of each score point, such as "4" (highest score), "3," "2", "1," and "0".

8. Choose only <u>one</u> of these "**Key Descriptors**" to indicate one that could be used to search the Pool for this item:

   - ❒ Vocabulary
   - ❒ Composition/Improvisation/Performance
   - ❒ History/Culture
   - ❒ Analysis/Evaluation
   - ❒ Careers/Interdisciplinary Connections

9. Check which **level of understanding** the item is assessing.

   ❒ Recall/Knowledge ❒ Higher (Comprehension, Application, Interpretation, Analysis, or Evaluation)

10. If this item incorporates stimulus material, complete the "Worksheet for Indicating a Stimulus Used in an Item;" the stimulus must be submitted along with the item. If no stimulus material, use the "Criteria for Creating and Evaluating Constructed-Response Items" checklist to review, evaluate, revise, and finalize the item drafted above.

WORKSHEET FOR INCLUDING A STIMULUS IN AN ITEM

USE THIS WORKSHEET FOR MULTIPLE-CHOICE OR CONSTRUCTED-RESPONSE ITEMS THAT INCORPORATE STIMULUS MATERIALS.

Stimulus Title _____
(a short name to identify this stimulus as distinct from others)

1. Type of Stimulus (choose only one from the following list):
   - ❒ Arts Reproduction (or other graphic)
   - ❒ Audio
   - ❒ Music Notation
   - ❒ Photograph
   - ❒ Video
   - ❒ Written/Quote

2. Brief Description of Stimulus:

_____

_____

_____

_____

3. Actual title of the stimulus: _____

4. Artist/author's full name: _____

5. Date completed/published (if known/applicable): _____

6. Use the "Checklist for Evaluating Stimulus Materials Used with Items" checklist to review, evaluate, revise, and finalize the item drafted above.

CHECKLIST FOR CREATING & EVALUATING MULTIPLE-CHOICE ITEMS

USE THE CHECKLIST BELOW IN DRAFTING, REVIEWING, EVALUATING, REVISING, AND FINALIZING MULTIPLE-CHOICE ITEMS.

SIGNIFICANCE OF THE ITEM

(Use a check mark to indicate "yes." All boxes must be marked "yes" before submitting item to Coach.)

- ❒ 1. The item relates directly to a national standard.
- ❒ 2. The item deals with information or a concept of importance, not with trivia.
- ❒ 3. The item is tightly focused so that it will be clear why a student gets it right or wrong.
- ❒ 4. If beyond recall or knowledge level, the item asks students to use higher-order thinking skills such as:
  - Comprehending an important concept or information
  - Applying a concept/skill to a different situation
  - Making an interpretation/drawing a conclusion
  - Comparing/contrasting concepts or information
  - Analyzing and/or evaluating a claim or generalization
- ❒ 5. The item addresses sensitivity issues and the concept of "universal design" for the broadest audience.

TECHNICAL QUALITIES OF THE ITEM

(Use a check mark to indicate "yes." All boxes must be marked "yes" before submitting item to Coach.)

- ❒ 1. The stem is a complete question or statement; the intent of the stem is clear without reading the options.
- ❒ 2. Information in the stem does not cue the key.
- ❒ 3. Negative stems are avoided.
- ❒ 4. There is one and only one clearly correct answer, and the key is indicated.
- ❒ 5. The position of the key is varied, while the order of the options is logical and appropriate.
- ❒ 6. All distractors are plausible, yet incorrect.
- ❒ 7. Overlapping options are avoided.
- ❒ 8. Options are parallel in concept, language structure, and appearance.
- ❒ 9. Options are of equal or nearly equal length.
- ❒ 10. Language usage and grammar in the stem and options are correct.
- ❒ 11. The stem and options fit correctly/appropriately with each other, grammatically and conceptually.
- ❒ 12. Complex formats in the stem, options, and overall item are avoided.
- ❒ 13. Repetitive language in the stem and the options is avoided.
- ❒ 14. Superfluous wording in the stem and options is avoided.
- ❒ 15. The item conforms to the SCASS Item Pool Style Guide for Multiple-Choice Selected-Response Items.
- ❒ 16. The use of "all of the above" and "none of the above" as options is avoided.
- ❒ 17. Items are independent to the extent possible; distractors are not used in more than one item.
- ❒ 18. Any stimulus materials fit with the stem, are of high quality, and are adequately identified.
- ❒ 19. Any stimulus materials conform to the specifications of the Item Pool.

CHECKLIST FOR CREATING & EVALUATING CONSTRUCTED-RESPONSE ITEMS

SIGNIFICANCE OF THE ITEM

(Use a check mark to indicate "yes." All boxes must be marked "yes" before submitting item to Coach.)

- ❒ 1. The item relates directly to a national standard.
- ❒ 2. The item deals with information or a concept of importance, not with trivia.
- ❒ 3. The item is tightly focused so that it will be clear why a student gets it right or wrong.
- ❒ 4. If beyond recall or knowledge level, the item asks students to use higher-order thinking skills such as:
  - Comprehending an important concept or information
  - Applying a concept/skill to a different situation
  - Making an interpretation/drawing a conclusion
  - Comparing/contrasting concepts or information
  - Analyzing and/or evaluating a claim or generalization
- ❒ 5. The item addresses sensitivity issues and the concept of "universal design" for the broadest audience.

TECHNICAL QUALITIES OF THE ITEM

(Use a check mark to indicate "yes." All boxes must be marked "yes" before submitting item to Coach.)

- ❒ 1. The intent of the stem and what the student is expected to do is clear.
- ❒ 2. Information in the stem does not inadvertently cue the response.
- ❒ 3. Negatives in the stem that would unnecessarily complicate understanding are avoided.
- ❒ 4. There is at least one plausible, accurate response to the stem.
- ❒ 5. The task demanded of the student is manageable within the timeframe.
- ❒ 6. The item is in "user-friendly" format.
- ❒ 7. Language usage and grammar in the stem are correct.
- ❒ 8. Superfluous wording in the stem is avoided.
- ❒ 9. Complex formats in the stem and overall item are avoided.
- ❒ 10. Scoring Criteria for 4 (highest), 3, 2, 1, and 0 are unambiguous and complete.
- ❒ 11. The stem and the scoring criteria fit correctly/appropriately with each other, grammatically and conceptually.
- ❒ 12. The item conforms to the SCASS Item Pool Style Guide for Constructed-Response Items.
- ❒ 13. Any stimulus materials fit with the stem, are of high quality, and are adequately identified.
- ❒ 14. Any stimulus materials conform to the specifications of the Item Pool.
- ❒ 15. A task description (for other than short constructed-response items) is provided.
- ❒ 16. Any special materials, equipment, or setup have been adequately described.
- ❒ 17. Directions and script for assessment administrators have been provided (for other than short constructed-response items).

COUNCIL OF CHIEF STATE SCHOOL OFFICERS

STATE COLLABORATIVE ON ASSESSMENT AND STUDENT STANDARDS

ARTS EDUCATION ASSESSMENT CONSORTIUM (SCASS/ARTS)

SCASS/ARTS WEBSITE ITEM POOL PROJECT

ASSIGNMENT OF COPYRIGHT

AGREEMENT entered into on this _____ day of _____, 20__, between _____ (Author), an individual, and the Council of Chief State School Officers (CCSSO), a nationwide, non-profit organization headquartered in Washington, D.C.

WHEREAS, Author has assisted in developing education assessment items for SCASS/Arts, a collaborative project of member states, and

WHEREAS, CCSSO is the legal entity that administers SCASS/Arts and wishes to obtain any copyright in said education assessment items;

NOW, THEREFORE, by this document and for good and valuable consideration given, Author does hereby assign and transfer to CCSSO, its successors and assigns, the entire right, title and interest in and to said copyright, and any renewal registrations and copyright applications relating thereto, and any renewals and extensions thereof, and in and to all works based upon, derived from, or incorporating the works covered by such copyrights, and in and to all income, royalties, claims and payments now or hereafter due or payable with respect thereto, and in and to all causes of action, either in law or equity for past, present, or future infringement based on said copyrights, and in and to all rights corresponding to the foregoing throughout the world.

Author agrees to execute all papers necessary to secure to CCSSO or its designee the rights herein assigned.

Signature _____

Title _____

COUNCIL OF CHIEF STATE SCHOOL OFFICERS

STATE COLLABORATIVE ON ASSESSMENT AND STUDENT STANDARDS

ARTS EDUCATION ASSESSMENT CONSORTIUM (SCASS/ARTS)

SCASS/ARTS WEBSITE ITEM POOL PROJECT

RELEASE FORM

SCASS/Arts is a consortium of state departments of education supported by the Council of Chief State School Officers (CCSSO), a nationwide, non-profit organization headquartered in Washington D.C. that provides assistance for "educational interests common to all of the states." SCASS/Arts Consortium was formed by CCSSO in 1992 for the purpose of addressing student assessment issues in arts education that are common to all states. The group is now developing a Website-based item pool of test questions (known as the **SCASS/Arts Website Item Pool Project**) to be shared by member states and used by educators in those states for their student assessment needs in arts education.

In the first phase of the project, selected-response (multiple-choice) arts education assessment items are developed by educators in the member states and/or refined from existing collections to meet the criteria of the Item Pool. In subsequent phases, this collection will be augmented by constructed-response items of all kinds including short-answer, essay, performance exercises, and processes for reviewing student work over time (portfolio).

The work you produce will be reviewed by different panels of content and assessment experts, refined if necessary, and, if it meets the established criteria for excellence, placed in the Web-accessed pool of items for the purposes stated above. By signing this document you grant CCSSO SCASS/Arts the right to use or not use, reproduce, perform, display, transmit, compile, collect, modify, and distribute any work completed by you for the stated purposes and goals. As with any work completed by educators from member states, the body of work will be copyrighted by CCSSO for the benefit of member states. If you would like to have your name associated with this body of work as a contributor, please check the appropriate box below. Attribution for individual items will not be provided.

This form provides documentation of your permission for us to perform these activities without profit to CCSSO or SCASS/Arts, and without compensation.

I, _____, have read this document and fully understand the contents thereof. I do hereby give SCASS/Arts and CCSSO right to use or not use, reproduce, perform, display, transmit, compile, collect, modify, and distribute any work created by me for the purposes and goals stated above.

Signature:_____ Date:_____

Check the box below to either grant or deny the SCASS/Arts Consortium permission to publish your name as a contributor to the **SCASS/Arts Website Item Pool Project**.

- ❑ Yes, you may use my name as a contributor to this body of work in any published venue.
- ❑ No, you may not use my name as a contributor to this body of work in any published venue.

COUNCIL OF CHIEF STATE SCHOOL OFFICERS

STATE COLLABORATIVE ON ASSESSMENT AND STUDENT STANDARDS

ARTS EDUCATION ASSESSMENT CONSORTIUM (SCASS/ARTS)

SCASS/ARTS WEBSITE ITEM POOL PROJECT

This glossary was developed by the members of SCASS/Arts Education Assessment Consortium. The terms below are words/phrases in general use in the field of assessment.

GLOSSARY OF ASSESSMENT TERMS

Accommodations

approved/standardized administrative or scoring adjustments (e.g., large print or Braille test booklets, individual or small group administrations, reading the test to the student) made for special populations taking standardized assessments

Accountability

having responsibility for; e.g., the public is holding educators and students accountable by demanding that schools demonstrate the impact and effectiveness of educational programs in order to justify the money invested in education. Accountability testing is supposed to provide achievement data, which will ultimately be used to improve the system.

Achievement test

a test designed to measure students' "school taught" learning, as opposed to their initial aptitude or intelligence.

Alternative assessment

assessments other than traditional multiple-choice tests; most often used to describe performance assessments or other assessments that provide more feedback about student learning than whether the answer is correct or incorrect.

Analytic scoring

A method of scoring performance assessments that yields multiple scores for the same task/performance. Performance is separated into major components, traits, or dimensions and each is independently scored. (e.g., a particular sample of a student's writing may be assessed as grammatically correct at the same time it is assessed as poorly organized.) Analytic scoring is especially effective as a diagnostic tool.

Anchor

(also called exemplars or benchmarks); a sample of student work (product or performance) used to illustrate each level of a scoring rubric; critical for training scorers of performances since it serves as a standard against which other student work is compared.

Aptitude test

a test which uses past learning and ability to predict what a person can do in the future; aptitude tests are depend more heavily on out-of-school experiences than in-school learning (Also see intelligence test.)

Assessment

The process of collecting and analyzing data for the purpose of evaluation. The assessment of student learning involves describing, collecting, recording, scoring, and interpreting information about performance. A complete assessment of student learning should include measures with a variety of formats as developmentally appropriate.

Authentic assessments

assessments that emulate the performance that would be required of the student in real-life situations.

Benchmarks

identifiable points on a continuum toward a goal or standard. The term may be used to describe content standards when interim targets (benchmarks) have been set by age, grade, or developmental level; the term is also used interchangeably with "anchor" papers or performances which illustrate points of progress on an assessment scale (i.e., student works which exemplify the different levels of a scoring rubric).

CIA

acronym for curriculum, instruction, and assessment

Cohort

a group of students whose progress is followed and measured at different points in time.

Competency test

a test intended to verify that a student has met standards (usually minimal) of skills and knowledge and therefore should be promoted, graduated, or perhaps deemed competent.

Constructed-response assessment

a form of assessment that calls for the student to generate the entire response to a question, rather than choosing an answer from a list (e.g., paper-and-pencil responses on essay or short answer tests or performances which may be drawn, danced, acted out, performed musically, or provided in any other way to exhibit particular skills or knowledge. (Also referred to as open-response and open-ended assessments.)

Context

the surrounding circumstances or environment in which an assessment takes place (e.g., embedded in the instruction or under standardized conditions [e.g., part of a large scale assessment])

Criteria

(sometimes used as synonym for traits or attributes); the rules or guidelines used for categorizing or judging; in arts assessment, the rules or guidelines used to judge the quality of a student's performance. (Also see rubric, scoring guide, and scoring criteria.)

Criterion-referenced assessment

an assessment designed to measure performance against a set of clearly defined criteria. Such assessments are used to identify student strengths and weaknesses with regard to specified knowledge and skills (which are the goals or standards of the instruction). Synonyms include: standard-based or -referenced, objective-referenced, content-referenced, domain-referenced, or universe-referenced.

Curricular alignment

the degree to which a curriculum's scope, sequence, and content match standards, instruction, assessment, or instructional resources.

Cut score

(also called performance standard) performance level or numerical score established by the assessment system to describe how well the student performed. The cut score can be manipulated to increase or decrease the number "passing" or "failing" a test. (Also see standard-setting.)

Descriptors

explanations that define the levels of scoring scales (Also see criteria.)

Dimension

specific traits, characteristics, or aspects of performance which are fairly independent of each other and can be scored separately (e.g., rhythm and melody can be scored separately for the same musical performance). Different scoring methods may be used for each dimension.

Disaggregate

(as in disaggregated data); pulling information apart (e.g., looking at the performance of various sub-groups instead of only the performance of the large group).

Educational outcome

an educational goal, expectation, or result that occurs at the end of an educational program or event (usually a culminating activity, product, or other measurable performance).

Enhanced/extended multiple-choice assessments

selected-response assessments with additional parts (for more points); this additional part often requires the students to justify their answers, show their work, or explain why they marked a particular option.

Essay test

a paper-and-pencil test that requires students to construct their entire brief or extensive responses to the question(s); should be limited to measuring higher levels of learning. (Also see constructed-response assessment.)

Extended response assessments

an essay question or performance assessment, which requires an elaborated or graphic response that expresses ideas and their interrelationships in a literate and organized manner

Evaluation

a judgment about the worth or quality of something. In education, data from tests, tasks, or performances are used to make judgments about the success of the student or program.

Formative Assessment

(Sometimes referred to as Assessment for Learning) A process used by teachers and students during instruction that provides feedback to adjust ongoing teaching and learning to improve students' achievement of intended instructional outcomes. Short-interval and usually classroom-based assessments that have immediate information for teachers and students to inform the instructional process and determine what comes next in the learning process.

Generalizability

the degree to which the performances measured by a set of assessment items/tasks are representative of the entire domain being assessed (E.g., is one performance assessment sufficient for drawing conclusions about a student's ability to critique works of art?); may also be an issue in drawing a sample of students from a population (i.e. the degree to which a sample of students is representative of the population from which it is drawn)

Grade equivalent

a score, available from some standardized tests, which describes the performance of students according to how it resembles the performance of students in various grades. A GE of 5.5 means that the student is performing like a student in the fifth month of the fifth grade.

Grading

a rating system for evaluating student work; grades are usually letters or numbers and their meaning varies widely across teachers, subjects, and systems.

High-stakes testing

any testing program for which the results have highly significant consequences for students, teachers, schools, and/or districts. These summative tests are frequently used as accountability devices to determine effectiveness or success.

Holistic method

a scoring method which assigns a single score based on an overall appraisal or impression of performance rather than analyzing the various dimensions separately. A holistic scoring rubric can be specifically linked to focused (written) or implied (general impression) criteria. Some forms of holistic assessment do not use written criteria at all but rely solely on anchor papers for training and scoring.

Intelligence tests

tests designed to measure general cognitive functioning; group or individually administered tests used to determine mental age as compared to chronological age (MA/CA x 100 = IQ [intelligence quotient]); i.e., the "average" IQ of the population is 100. Some intelligence tests do not calculate mental age but compare an individual's performance to the performance of a norm group at various developmental levels, generating verbal and performance scores with a mean or "average" score of 100.

Item analysis

a statistical analysis of the items on a selected-response test to determine the relationship of the item to the test's validity and reliability as a whole. The number and nature of the students selecting each option are analyzed.

Matrix sampling

a process used to estimate the performance of large groups through testing a representative sample of the students. Each student in the sample may be given only a small segment of the total assessment.

Mean

the arithmetic average of a group of scores; one of three measures of central tendency, a way to describe a group of scores with a single number.

Median

a measure of central tendency, which identifies the point on the scale that separates a group of scores so that there is an equal number of scores above and below it.

Metacognition

the ability to think about one's own thinking; the knowledge that individuals have of their own thinking processes and strategies and their ability to monitor and regulate those processes.

Multiple-choice test

a test consisting of items (questions or incomplete statements) followed by a list of choices from which students have to select the correct or best response.

Multiple measures

the use of a variety of assessments to evaluate performance in a subject area (e.g., using multiple-choice items, short answer questions, and performance tasks to assess student achievement in a subject); the use of multiple measures is advocated to obtain a fair and comprehensive measurement of performance

Mode

a measure of central tendency which identifies the most frequent score in a group of scores (e.g., in the group of scores: 1, 2, 8, 9,9,10, the mode is 9).

Norm

the midpoint or "average" score for the group of students to which a norm-referenced test was initially administered (the norm group). By design, 50% of the students score below and 50% above this score.

Norm group

a group of students that is first administered a standardized norm-referenced test by its developers in order to establish scores for interpreting the performance of future test-takers.

Norm-referenced test

a standardized test which compares the performance of students to an original group that took the test (the norm group); results usually reported in terms of percentile scores (e.g., a score of 90 means that the student did better than 90% of the norm group).

Normal curve equivalent (NCE)

a normalized standard score used to compare scores across tests with different scales and/or between students on the same test (since arithmetic manipulations should not use percentiles); it has a mean of 50, a standard deviation of 21.06 and is often required for reporting by federal funding agencies such as Title I.

Open-ended assessments

constructed assessments (frequently tasks or problems) that require students to generate a solution to a problem for which there is no single correct answer (e.g., create a drawing that uses symbols of the Renaissance)

Open-response assessments

constructed assessments (ones for which students must construct the entire answer and show their work) that have a single correct answer but multiple methods of solution possible

Percentile

a statistic provided by standardized norm-referenced tests which describes the performance of a student as compared to that of the norm group. The range is 1to 99 with 50 denoting average performance. A student scoring at the 65th percentile performed better than, or as well as, 65% of the norm group.

Performance assessment

a task/event/performance designed to measure a student's ability to directly demonstrate particular knowledge and skills. E.g., a student may be asked to demonstrate some physical or artistic achievement: play a musical instrument, create or critique a work of art, or improvise a dance or a scene. These kinds of assessments (e.g., tasks, projects, portfolios, etc.) are scored using **rubrics**: established criteria for acceptable performance.

Performance-based instruction

See standards-based instruction

Portfolio

a purposeful collection of student work across time which exhibits a student's efforts, progress, or level of proficiency. Examples of types of portfolios include: showcase (best work), instructional, assessment (used to evaluate the student, and process or project (shows all phases in the development of a product or performance).

Primary trait scoring

A type of rubric scoring constructed to assess a specific trait, skill or format or the impact on a designated audience. (Also see analytic scoring.)

Project

a type of performance assessment which is complex, usually requiring more than one type of activity, process, or product for completion.

Quartile

a way of describing the position of a score on a norm-referenced test, e.g., the score falls in one of four groups: 0-25th percentile, 26-50th percentile, etc.

Quintile

a way of describing the position of a score on a norm-referenced test, e.g., the score falls in one of five groups: 0-20th percentile, 21-40th percentile, etc.

Range

the most rudimentary method of describing how much a group of scores vary; range is determined by subtracting the lowest from the highest score in the group

Rating scale

a scale used to evaluate student learning using a gradation of numbers or labels; a Likert rating scale is frequently used to measure attitudes or perceptions

Reliability

a measure of the consistency of an assessment across time, judges and subparts of the assessment (assuming no real change in what is being measured).

Rating scale

a scale used to evaluate student learning using numbers or labels (e.g, a Likert scale).

Rubric

(sometime referred to as a scoring guide or scoring criteria) an established, ordered set of criteria for judging student performance/products; it includes performance descriptors of student work at various levels of achievement.

Sampling

A way to get information about a large group by examining a smaller representative number of the group (the sample).

Scale score

a score indicating an individual's performance on a standardized test, which allows comparisons across sub-groups and time. (E.g., one could use scale scores to compare test results among classes, schools, and districts; or across grades from year to year.)

Scaffolded assessments

a set of context-dependent assessments, which are sequenced to measure ascending levels of learning; this set usually contains a variety of item formats (from multiple-choice to performance tasks) about a single stimulus (e.g., a specific set of materials: a particular situation, scenario, problem, or event). Since these kinds of assessments can measure a variety of kinds of learning, they provide the opportunity for diagnosis of instruction and identification of student strengths and weaknesses.

Scoring criteria

the rules or guidelines used to assign a score (a number or a label) indicating the quality of a performance; in the analytic scoring of a performance, different rules may be applied to different dimensions or traits of the performance.

Scoring guide

directions for scoring and/or interpreting scores; the guide may include general instructions for raters, training notes, rating scales, rubric, and student work.

Selected-response items

a kind of test item for which students have to select the best or correct answer from a list of options (multiple-choice, etc.) or indicate the truth or falsity of a statement.

Self-assessment

collecting data about one's own performance for the purpose of evaluating it. Self-evaluation may include the comparison of one's own performance against established criteria, change in performance over time, and/or a description of current performance.

Standard

Three types of educational standards are frequently used in education today:

Content standards specify what students should know and be able to do in a specific content area---the essential knowledge, skills, processes, and procedures students must learn and be able to demonstrate. They answer the question: "What should be learned in this subject?" Student standards have been developed for periods of time ranging from individual grade levels to lifelong learning.

Performance standards specify the degree or quality of learning students are expected to demonstrate in the subject. They answer the question: "How good is good enough?" The national standards for the arts use the term "achievement standards" to avoid confusion between arts performance and performance assessment. (Some states refer to established levels of proficiency instead of performance standards.)

Opportunity-to-learn standards specify what schools must provide to enable students to meet content and performance standards.

Standard deviation

a measure of the variability of a group of scores. When the standard deviation is high, students are performing very differently from each other; if it is low, students are performing similarly to one another.

Standard error of measurement

a statistic used to indicate the consistency and reliability of a measurement instrument; a large standard error of measurement indicates that we have less confidence in the obtained score

Standards-based instruction

instruction designed, taught, and assessed using student standards (achievement targets)

Stanine

A standard 9-point scale used to report the results of norm-referenced tests in order to allow comparison of scores across students, schools, districts, tests, grades, etc. The mean is 5 and the standard deviation approximately 2. Stanines of 1-3 are considered below average; 4-6 average; and 7-9 above average.

Standardized Test

A test administered to a group of persons under the same specific conditions so student results can be fairly compared.

Summative Assessment

The effort to summarize student learning at a particular point in time such as the end of a chapter, unit, grading period, semester, year, or end of course.

Test

A sample of behavior or performance administered in order to provide a basis for inferences about a larger subject area or domain of study. E.g., a teacher may administer a 30-minute test to provide evidence of the student's learning for the last two weeks or for a particular unit of instruction. The test may be norm- or criterion-referenced, traditional (e.g., multiple-choice, short answer, essay, etc.), or performance-based. A *teacher-made test* is one prepared and administered by the teacher, usually for use in the classroom.

Validity

A characteristic of a measure which refers to its ability to measure what it is intended to measure AND do so **reliably** (i.e., measures consistently across time, judges, and sub-parts). A valid measure is both accurate and consistent; e.g., a bathroom scale may record 100 pounds every time a woman gets on it, but if she actually weighs 120, the scale is reliable but not valid. Types of validity include:

Content validity—The assessment has content validity if it measures the content or area it intends to measure.

Concurrent validity—The assessment has concurrent validity if it is correlated with other measures of that particular content or area.

Predictive validity—The assessment has predictive validity if it predicts later actual performance of the individual in that subject or area. Predictive validity is related to generalizability.